# Capítulo 12

Uso de la Distribución Bernoulli Multivariada en salud bucal

Ramón Álvarez-Vaz y Fernando Massa

Universidad de la República,
Facultad de Ciencias Económicas y de Administración,
Departamento de Métodos Cuantitativos,
Instituto de Estadística,
Eduardo Acevedo 1139,
C.P. 11200, Montevideo, Uruguay,
ramon@iesta.edu.uy,fmassa@iesta.edu.uy,
unidad.biometria.iesta@gmail.com

**Resumen.** En general, en muy variadas disciplinas como la Economía, el Marketing, la Epidemiología, se dan situaciones donde la matriz de datos de la que se dispone está formada por datos binarios (unos y ceros) que surgen de trabajar con varias variables aleatorias resultantes de un experimento con 2 resultados posibles en cada caso. El interés se centra entonces, generalmente, en analizar y dar cuenta de las relaciones que se dan entre variables a través de la distribución Bernoulli Multivariada **(BM)**. Esta distribución puede ser caracterizada por un vector de *intensidades* y una matriz de *asociaciones* entre las variables binarias, que se pueden interpretar y asimilar como los parámetros de un modelo de regresión, por lo cual es importante entonces ver como queda parametrizado este modelo probabilístico y como puede ser estimado.

Se presenta luego a modo de ejemplo una aplicación en salud bucal para evaluar la enfermedad periodontal en la población adulta uruguaya. Los datos surgen del primer relevamiento nacional de salud bucal, llevado a cabo durante los años 2011 y 2012 en diversos departamentos de Uruguay, donde fueron encuestadas personas de 3 grupos etarios (jóvenes, adultos y adultos mayores), a los que se les evalúa presencia de enfermedad periodontal, evaluada como atributos binarios en 6 sextantes de la boca, por lo cual se tienen 6 variables binarias.

**Abstract.** In general in very varied disciplines such as Economics, Marketing, and Epidemiology there are situations where the available data matrix is formed by binary data (ones and zeros) that arise from working with several random variables resulting from an experiment with 2 possible results in each case. The interest is then generally focused on analyzing and accounting for the relationships that occur between variables through the Multivariate Bernoulli (MB) distribution presented

in this work. This distribution can be characterized by a vector of intensities and a matrix of associations between binary variables, which can be interpreted and assimilated as the parameters of a regression model, so it is important to see how it is parameterized this probabilistic model and how it can be estimated. An oral health application is then presented as an example to evaluate periodontal disease in the Uruguayan adult population measured as binary attributes in 6 sextants of the mouth, for which there are 6 binary variables.

*Palabras clave:* asociación, distribución Bernoulli Multivariada, enfermedad periodontal, intensidad, variable latente.

## 12.1. Introducción

En este documento se presenta y caracteriza una distribución de probabilidad multivariada que solo puede adoptar los valores cero o uno y que se denomina *Bernoulli Multivariada* (BM). Esta distribución equivale a considerar los vértices de un hipercubo en $\mathbb{R}^k$, cuyas coordenadas son los valores 0 y 1. Una de las primeras aproximaciones a la temática se puede encontrar en [9] donde se plantea la distribución de Bernoulli bivariada.

En primera instancia, la distribución BM podría definirse como el producto de $k$ distribuciones marginales cada una acorde al modelo Bernoulli [12], sin embargo dicha parametrización solo contempla el caso en el que las variables en cuestión son independientes. Es por esto que aquí se opta por una formulación donde se incluye la opción de modelar las asociaciones entre las variables. Para ello, se siguen las ideas expuestas en [4]. Pese a que la naturaleza categórica de las variables permite pensar en asociaciones entre dos, tres o mas de éllas simultáneamente, se toma la decisión de contemplar solamente las asociaciones "dos a dos" a modo de construir modelos mas parsimoniosos.

Sin embargo, la metodología aquí propuesta puede extenderse para tener en cuenta asociaciones de orden superior. El método empleado en este trabajo difiere de la parametrización basada en la dicotomización de la distribución Gaussiana multivariada [3] [12] debido a que, a diferencia de esta, no asume la existencia de variables latentes, lo cual supone una ventaja en cuanto a la simplicidad del modelo probabilístico.

El documento se estructura de la siguiente manera. En la sección 12.2 se considera la construcción de la distribución, comenzando desde el caso univariado, pasando por el bivariado y llegando finalmente al modelo general, presentando las principales propiedades de cada caso. En la sección 12.3 se presenta una aplicación en salud oral de esta metodología. Se plantean algunos estadísticos para explorar la independencia o asociación entre las variables. En la sección 12.4 se plantea un resumen de los resultados encontrados y posibles caminos por donde seguir.

## 12.2. Modelo probabilístico

A continuación se plantea la distribución BM comenzando como una reparametrización de la distribución de Bernoulli, para luego extenderla al caso bivariado y finalmente al caso general. En cada etapa se exploran las principales características de la función de masa de probabilidad.

### 12.2.1. Caso univariado

La distribución Bernoulli es utilizada para modelar las variables aleatorias resultantes de un experimento binario (considerando $Rec(X) = \{0, 1\}$) mediante un único parámetro, el cual se interpreta como la probabilidad de obtener un éxito en dicho experimento. La función de cuantía es la siguiente:

$$P(X = x) = p^x(1-p)^{1-x},\ x \in \{0, 1\}. \tag{12.1}$$

La variable aleatoria definida de esta manera tiene esperanza $p$ y varianza $p(1-p)$. También, es sencillo apreciar que esta función de cuantía puede expresarse como un miembro de la familia exponencial.

$$P(X = x) = e^{xlog(\frac{p}{(1-p)})+log(1-p))},\ x \in \{0, 1\}. \tag{12.2}$$

De esta manera surge que el "parámetro natural" de esta distribución es el logaritmo del *odd*. Tras llevar a cabo el cambio de variable $\phi_1 = \frac{p}{(1-p)}$, se llega a la siguiente parametrización:

$$P(X = x) = \phi_0\phi_1^x,\ x \in \{0, 1\}, \tag{12.3}$$

donde $\phi_1$ representa el *odd* de éxito y $\phi_0$ es una constante que normaliza la distribución y que se interpreta como la probabilidad de obtener un fracaso. En el caso univariado, esta constante es $\phi_0 = \frac{1}{1+\phi_1}$. Las nuevas expresiones para la esperanza y varianza de la distribución son $E(X) = \frac{\phi_1}{1+\phi_1}$ y $Var(X) = \frac{\phi_1}{(1+\phi_1)^2}$. Pese a que, en un principio, esta reparametrización solo parece complicar la caracterización de la distribución, en dimensiones superiores probará ser de gran utilidad ya que proporcionará gran flexibilidad para incluir las asociaciones entre variables.

### 12.2.2. Caso bivariado

En el caso bivariado, si las variables $X_1$ y $X_2$ son independientes, su cuantía conjunta podría definirse de la siguiente manera:

$$P(X_1 = x_1, X_2 = x_2) = p_1^{x_1}(1-p_1)^{1-x_1}p_2^{x_2}(1-p_2)^{1-x_2}, X_1, X_2 \in \{0, 1\}^2 \tag{12.4}$$

Luego de realizar el mismo cambio de variable sugerido en el apartado anterior, la cuantía conjunta se expresa de la siguiente manera:

$$P(X_1 = x_1, X_2 = x_2) = \phi_0 \phi_1^{x_1} \phi_2^{x_2}, X_1, X_2 \in \{0,1\}^2 \tag{12.5}$$

En este caso, la constante de normalización $\phi_0$ equivale a $\frac{1}{1+\phi_1+\phi_2+\phi_1\phi_2}$ y se interpreta como la probabilidad de obtener un fracaso en ambas variables. El siguiente paso en la construcción de la distribución BM es el de incluir en la ecuación (12.5) la asociación entre $X_1$ y $X_2$. Para ello se introducirá un nuevo parámetro $\alpha_{12}$ de la siguiente manera:

$$\begin{array}{l} P(X_1 = x_1, X_2 = x_2) = \phi_0 \phi_1^{x_1} \phi_2^{x_2} \alpha_{12}^{x_1 x_2}, X_1, X_2 \in \{0,1\}^2 \\ \phi_0 = \frac{1}{1+\phi_1+\phi_2+\phi_1\phi_2\alpha_{12}} \end{array} \tag{12.6}$$

Tras la modificación propuesta, $\phi_0$ continúa siendo la probabilidad de obtener dos fracasos.

En cuanto al parámetro $\alpha_{12}$, es sencillo demostrar que equivale al *odds ratio* entre $X_1$ y $X_2$. Sin embargo, la interpretación de $\phi_1$ y $\phi_2$ cambia ligeramente, ya que en este caso pasan a ser los *odds* de éxito de cada variable *condicional* a que la otra variable valga cero. Ya en presencia de ambos tipos de parámetros, nos referiremos al conjunto de valores $\phi_i$ como *intensidades* o *fuerzas* y al conjunto de valores $\alpha_{ij}$ como *asociaciones*.

El siguiente paso es definir las distribuciones marginales y condicionales de cada variable. En cuanto a las marginales, se puede demostrar que son Bernoulli con la siguiente función de cuantía:

$$\begin{array}{l} P(X_1 = x) = \phi_0^* \phi_1^{*x}, X_1, X_2 \in \{0,1\} \\ \phi_1^* = \phi_1(1 + \phi_2\alpha_{12}), \\ \phi_0^* = \frac{1}{1+\phi_1^*} \end{array} \tag{12.7}$$

La distribución marginal de $X_2$ es análoga. En cuanto a las distribuciones condicionales, se puede demostrar que éstas también son Bernoulli:

$$\begin{array}{l} P(X_1 = i|X_2 = j) = \frac{P(X_1=i,X_2=j)}{P(X_2=j)} = \frac{\phi_0\phi_1^i\phi_2^j\alpha_{12}^{ij}}{\phi_0\phi_2^j(1+\phi_1\alpha_{12}^j)} = \frac{\phi_1^i\alpha_{12}^{ij}}{(1+\phi_1\alpha_{12}^j)} = \phi_{0|j}\phi_{1|j}^i, (i,j) \in \{0,1\}^2, \\ \phi_{1|j} = \phi_1\alpha_{12}^j, \\ \phi_{0|j} = \frac{1}{1+\phi_1\alpha_{12}^j}. \end{array} \tag{12.8}$$

Vale la pena mencionar que al fijar $j = 0$ se obtiene el caso particular de donde surge la interpretación de $\phi_1$ como *odd* condicional. La cuantía condicional de $X_2$ se obtiene de la misma manera.

### 12.2.3. Caso general

La función de cuantía del vector $X = (X_1, X_2, \ldots, X_k)$ en el caso de $k$ variables binarias posiblemente asociadas entre sí es la siguiente:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \phi_0 \prod_{i=1}^{k} \phi_i^{x_i} \prod_{j=i+1}^{k} \alpha_{ij}^{x_i x_j}, x_1, x_2, ..., x_k \in \{0,1\}^k \tag{12.9}$$

En este caso, la especificación de $\phi_0$ se vuelve un poco mas compleja y para ello se define la matriz de configuraciones $H$. Esta matriz, que consta de $\frac{k(k+1)}{2}$ columnas y $2^k$ filas, contiene cada una de las posibles configuraciones del vector aleatorio $X$ en las primeras $k$ columnas y los productos de estas coordenadas en las siguientes $\frac{k(k-1)}{2}$. Adicionalmente se define el vector $\gamma$, el cual contiene los $\frac{k(k+1)}{2}$ parámetros del modelo. Se trabaja entonces con $\Gamma = (log\phi_1, log\phi_2, \ldots, log\alpha_{k-1,k})$. De esta manera, se reescribe la cuantía en función de los elementos de $\Gamma$.

$$\begin{aligned} P(\underline{X} = \underline{x}) &= \phi_0 \prod_{i=1}^{k} \phi_i^{x_i} \prod_{j=i+1}^{k} \alpha_{ij}^{x_i x_j}, \underline{X} \in \{0,1\}^k \\ &= exp(log(\phi_0 \prod_{i=1}^{k} \phi_i^{x_i} \prod_{j=i+1}^{k} \alpha_{ij}^{x_i x_j})) \\ &= \phi_0 exp(\sum x_i log\phi_i + \sum x_i x_j log\alpha_{ij}). \end{aligned} \tag{12.10}$$

Y, al sumar todos los elementos de la cuantía:

$$\begin{aligned} 1 &= \sum_{x \in H} P(\underline{X} = \underline{x}) = \phi_0 \sum_{x \in H} exp(\sum x_i log\phi_i + \sum x_i x_j log\alpha_{ij}) \\ \Rightarrow \phi_0 &= \frac{1}{\sum_{x \in H} exp(\sum x_i log\phi_i + \sum x_i x_j log\alpha_{ij})} \\ \Rightarrow \phi_0 &= \frac{1}{\mathbf{1} e^{H\phi}} \end{aligned} \tag{12.11}$$

A modo de ejemplo se presenta el caso particular $k = 2$. En dicho caso $\Gamma$ y H adoptan la siguiente forma:

$$\Gamma = \begin{vmatrix} log\phi_1 \\ log\phi_2 \\ log\alpha_{12} \end{vmatrix} \quad H = \begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{vmatrix}$$

De esta manera:

$$\begin{aligned}\phi_0 &= \frac{1}{\mathbf{1}e^{H\Gamma}} \\ &= \frac{1}{e^{<(0,0,0)\Gamma>}+e^{<(1,0,0)\Gamma>}+e^{<(0,1,0)\Gamma>}+e^{<(1,1,1)\Gamma>}} \\ &= \frac{1}{e^0+e^{log\phi_1}+e^{log\phi_2}+e^{log\phi_1+log\phi_2+log\alpha_{12}}} \\ &= \frac{1}{1+\phi_1+\phi_2+\phi_1\phi_2\alpha_{12}},\end{aligned}$$

tal como se vio en la ecuación (12.6).

En cuanto a la interpretación de los parámetros, $\phi_0$ continúa interpretándose como la probabilidad de obtener el valor cero en todas las variables. En cuanto a los parámetros $\phi_i$ y $\alpha_{ij}$, éstos se interpretan como los *odds* y *odds ratio* condicionales a que el resto de las variables sean cero. Pese a que sería deseable que la interpretación de dichos coeficientes no fuese parcial, es fi¿$\frac{1}{2}$cil construir estimadores incondicionales a partir de los elementos del vector $\Gamma$.

El siguiente paso es definir las distribuciones condicionales y marginales de subconjuntos del vector $X$. Sin pérdida de generalidad se asumirá que se quiere obtener la distribución marginal del vector $X^M = (X_1, X_2, \ldots, X_M)$, la cual se obtendrá sumando sobre los $2^{k-M}$ valores posibles del vector $X^m = (X_{M+1}, \ldots, X_k)$, donde $m = k - M$.

$$\begin{aligned}P(X^M = x) &= \sum_{X_{M+1}=0}^{X_{M+1}=1} \cdots \sum_{X_k=0}^{X_k=1} \phi_0 \prod_{i=1}^{p} \phi_i^{x_i} \prod_{j=i+1}^{p} \alpha_{ij}^{x_i x_j} \\ &= \phi_0^* \prod_{i=1}^{M} \phi_i^{x_i} \prod_{j=i+1}^{M} \alpha_{ij}^{x_i x_j} F(x, \phi^m, \phi^{Mm}), x \in \{0,1\}^M \\ &= \phi_0^* \prod_{i=1}^{M} \phi_i^{*x_i} \prod_{j=i+1}^{M} \alpha_{ij}^{*\, x_i x_j}.\end{aligned}$$

De aquí se puede concluir que todas las distribuciones marginales también pertenecen a la familia de distribuciones BM. En cuanto a $F(x, \phi^m)$, es una función que involucra a los elementos de $x$, a las intensidades $(\gamma^{(m)})$ correspondientes a las variables sobre las cuales se suma y a las intensidades $(\phi^M)$ que "vinculan" los elementos de $X^M$ y $X^m$. Para la construcción de los parámetros marginales se utiliza el resultado anterior conjuntamente con la definiciones de *odd* y *odds ratio* marginales. La definición de las intensidades marginales $\phi_i^*$ en (12.13) es la siguiente:

$$\phi_i^* = \phi_i \widetilde{\gamma}_i^{(k-M)}, \tag{12.14}$$

donde $\widetilde{\phi}_i^{(m)} = \frac{e^{H\phi^{(m)}}}{\mathbf{1}e^{H\phi^m}}\phi^{M(m)}$. La interpretación de estas intensidades marginales corresponde a una corrección de las intensidades originales, donde dicha corrección se construye como un promedio ponderado de las asociaciones $(\alpha^{M(m)})$ entre $X_i$ y las variables contenidas en $X^k$, con ponderadores dados por las intensidades y asociaciones $(\alpha^m)$ de las variables sobre las cuales se sumó.

El caso de las asociaciones marginales en la ecuación (12.13) es similar:

$$\alpha_{ij}^{*} = \alpha_{ij}\phi_i\phi_j \frac{\widetilde{\gamma}_{ij}^{(m)}}{\widetilde{\gamma}_{i}^{(m)}\widetilde{\gamma}_{j}^{(m)}}. \tag{12.15}$$

Las distribuciones condicionales son mas sencillas y se construyen a partir de la siguiente relación:

$$P(X^M = x^M | X^{(m)} = x^m) = \frac{P(X^M = x^M, X^m = x^m)}{P(X^m = x^m)}, X^M \in \{0,1\}^M \tag{12.16}$$

donde el numerador no es otra cosa que la cuantía que ya se definió en (12.9) y el denominador corresponde a la marginal del vector $X_k$ que se acaba de presentar. Finalmente la cuantía condicional es la siguiente:

$$\begin{aligned} &P(X^M = x^M | X^m = x^m) = \phi_{0|j} \prod_{i=1}^{M} \phi_{i|m}^{x_i} \prod_{j=i+1}^{M} \alpha_{ij}^{x_i x_j}, X^M \in \{0,1\}^M, X^m \in \{0,1\}^{k-M} \\ &\phi_{0|j} = \frac{1}{\mathbf{1}e^{H\phi_{M|m}}}, \\ &\phi_{i|m} = \phi_i \prod_{x_j \in m} \alpha_{ij}^{x_j}. \end{aligned} \tag{12.17}$$

Hay que tener en cuenta cómo el proceso de condicionar en los valores de las variables contenidas en $X^m$ solo afecta las intensidades y no las asociaciones.

### 12.2.4. Estimación

Dado que la función de verosimilitud es no lineal en los parámetros, se opta por realizar la estimación de los parámetros del modelo BM mediante técnicas de optimización numérica. Para ello, se define la función de log-verosimilitud de una muestra de $n$ observaciones como:

$$\ell(\underline{x}|\underline{\phi}) = n\log(\phi_0) + \sum_{j=1}^{k} S_j \log\phi_i + \sum_{j=1}^{k} S_{jk}\log\alpha_{ik}, \underline{x} \in \{0,1\}^k, \tag{12.18}$$

donde $S_j = \sum_{i=1}^{n} x_{ij}$ y $S_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik}$.

La maximización de esta función se lleva a cabo por algunos de los métodos iterativos comunmente utilizados. La mayoría de los mismos requiere del gradiente (o *score*) y la matriz Hessiana de la ecuación (12.18). Los elementos del primero (al que denotamos como $U(\underline{\phi})$) tienen la siguiente forma:

$$\begin{aligned} U_j(\underline{\gamma}) &= \frac{\partial \ell(\underline{x}|\underline{\phi})}{\partial \phi_j} = \frac{S_j}{\phi_j} - n\frac{\mathbf{1}e^{H_{(j)}\Gamma}}{\mathbf{1}e^{H\Gamma}}, \\ U_{jk}(\underline{\gamma}) &= \frac{\partial \ell(\underline{x}|\underline{\phi})}{\partial \alpha_{jk}} = \frac{S_{jk}}{\alpha_{jk}} - n\frac{\mathbf{1}e^{H_{(jk)}\Gamma}}{\mathbf{1}e^{H\Gamma}}, \end{aligned} \tag{12.19}$$

donde $H_{(j)}$ es la matriz compuesta por las filas de $H$ que contienen unos en la j-ésima columna (correspondiente a $\phi_j$), luego esta columna es reemplazada por un vector de ceros. El caso de $H_{(jk)}$ es análogo al anterior pero con la columna correspondiente a $\alpha_{jk}$ reemplazada por un vector de ceros.

## 12.3. Una aplicación a la salud oral

Una posible aplicación de esta distribución es en al análisis de la enfermedad periodontal. La enfermedad periodontal, es una de las enfermedades más prevalentes en Odontología, teniendo un peso muy importante en la carga mundial de enfermedades no trasmisibles (ENT), que afectan al 40 % de la población mundial [8]. Desde el punto de vista de la salud colectiva el estudio de su distribución, explicación, prevención y tratamiento debe abordarse integralmente y considerase en el contexto de la salud general de los colectivos humanos. Desde el punto de vista biológico, la enfermedad periodontal está asociada al biofilm, matriz de microorganismos (incluidos los patógenos en una baja proporción) adheridos a la superficie del diente que en condiciones normales, se encuentran en armonía con el huésped sano. Los signos asociados con esta patología son sangrado gingival, sarro, bolsa patolígica, pérdida de inserción de los tejidos periodontales, pérdida ósea y movilidad dentaria. Los índices que pretenden dar cuenta de la enfermedad periodontal tienen limitaciones derivadas del número de signos involucrados así como de los instrumentos utilizados y la subjetividad del observador. A nivel internacional se habla de enfermedad periodontal cuando existen bolsas periodontales iguales o mayores a 4 mm, la que se mide a través del índice CPI.

### 12.3.1. Datos de sangrado

A continuación se presenta la aplicación de la distribución BM para el análisis del sangrado peridontal que es uno de los componentes de la enfermedad periodontal.

Se trabaja con los datos provenientes del estudio sobre personas que demandan atención en la Facultad de Odontología de la Universidad de la República, Uruguay y que son evaluados por los odontólogos del Servicio de registros de la Facultad. Se aplica una muestra de 602 personas que consultan en el período que corresponde a mayo 2015-junio 2016, los que se seleccionan mediante muestreo sistemático, a los que se les aplica un cuestionario sociodemográfico y un examen completo de la boca, en donde se evalúa el estado de las piezas dentales y de la mucosa.

Vemos como ejemplo 6 registros de la tabla de datos que muestran el estado en términos de sangrado para las diferentes piezas que componen cada sextante, tal como aparece en el Cuadro 12.1

En la Figura 12.1 puede verse que hay sextantes vinculados al maxilar superior (sextantes 1, 2 y 3) e inferior (sextantes 4, 5 y 6) y a su vez si están en la parte derecha (sextantes 1 y 6) o izquierda (sextantes 3 y 4) de la boca.

| paciente | s11 | s31 | s1617 | s2627 | s3637 | s4647 |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 0 | 1 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| 550 | 0 | 0 | 1 | 1 | 0 | 0 |

Cuadro 12.1: Ejemplo de Presencia de sangrado en 6 personas

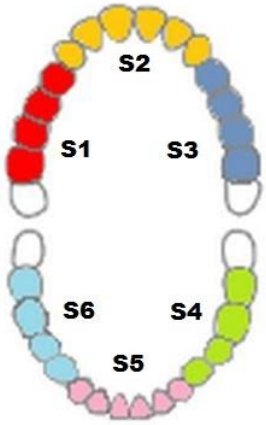


Figura 12.1: Distribución de los sextantes en la boca

| piezas | sextante | presencia | ausencia | % |
|---|---|---|---|---|
| piezas 16 y 17 | S1 | 167 | 435 | 27,7 |
| pieza 11 | S2 | 129 | 473 | 21,4 |
| piezas 26 y 27 | S3 | 174 | 428 | 28,9 |
| pieza 31 | S4 | 161 | 441 | 26,7 |
| piezas 36 y 37 | S5 | 119 | 483 | 19,8 |
| piezas 46 7 47 | S6 | 134 | 468 | 22,2 |

Cuadro 12.2: Presencia de sangrado por sextantes

En un análisis preliminar de estos datos se observó que gran parte de la muestra tiene esta patología. Se consideró "sano" a un individuo con sus 6 sextantes sanos. Estos constituyen apenas el 43.8 % de los datos, por lo tanto conformaron un perfil claro de individuos los cuales se dejaron de lado para trabajar sobre el resto, de modo de poder determinar distintos perfiles de carga de enfermedad.

| Cantidad de sextantes con sangrado | Frecuencia |
|---|---|
| 0 | 264 |
| 1 | 121 |
| 2 | 61 |
| 3 | 68 |
| 5 | 33 |
| 5 | 25 |
| 6 | 30 |
| Total | 602 |

Cuadro 12.3: Distribución de Cantidad de sextantes con sangrado

A partir de estos datos se van a ajustar modelos donde se supone que no hay restricciones entre las relaciones de las 6 variables y luego modelos donde hay independencia local y homogeneidad local de las asociaciones.

### 12.3.2. Modelo ajustado

Las subrutinas de cálculos fueron desarrolladas en el sistema R [11] usando, para la optimización los algoritmos de optimización no lineal implementados en la librería *nloptr* [5] y que aparecen comentados por Ypma en el reporte técnico [14].

A continuación se muestran las subrutinas de estimación creadas en R especialmente con los resultados de las estimaciones puntuales y por intervalo, por ejemplo para el modelo simple (sin restricciones).

```
y<-datos[,c(15:20)]
modelo1<-estim(y)
L0<-c(modelo1$intensidades,modelo1$asociaciones)
int.conf(modelo1,0.05)
repar(modelo1$intensidades,modelo1$asociaciones)
```

Vemos entonces los valores estimados $\hat{\phi}_i$ y $\hat{\alpha}_{ij}$ que devuelve la función *estim*. Por otra parte, para una mejor interpretación de lo resultados, se reparametrizan los $\hat{\phi}_i$ y los $\hat{\alpha}_{ij}$ para ser presentados como *proporciones*,*odds* y *OR*

```
> modelo1[1:2]

$intensidades
[1] 0.097 0.066 0.124 0.076 0.130 0.045

$asociaciones
```

```
 [1] 2.678 4.675 1.585 2.423 1.871 1.880 1.777 2.859
 2.513 2.140 1.271 2.665 2.182 4.970 1.916

repar(modelo1$intensidades,modelo1$asociaciones)
$proporciones
[1] 0.214 0.268 0.277 0.289 0.198 0.223

$odds
[1] 0.273 0.365 0.384 0.407 0.2464 0.286

$OR
 [,1]  [,2]  [,3]  [,4]   [,5]   [,6]
[1,]   Inf 6.535 7.237 5.7414 7.875  5.949
[2,] 6.535   Inf 5.532 3.9374 5.864  5.396
[3,] 7.237 5.532   Inf 9.0184 6.901  5.426
[4,] 5.741 3.937 9.018   Inf  7.662  5.906
[5,] 7.875 5.864 6.901 7.6617    Inf 11.56
[6,] 5.949 5.396 5.426 5.9057 11.56    Inf
```

| intensidades | | | | | |
|---|---|---|---|---|---|
| S1 | S2 | S3 | S4 | S5 | S6 |
| 0.097 | 0.066 | 0.124 | 0.076 | 0.130 | 0.045 |
| asociaciones | | | | | |
| - | 2.67 | 4.67 | 1.58 | 2.42 | 1.87 |
| | - | 1.88 | 1.77 | 2.85 | 2.51 |
| | | - | 2.14 | 1.27 | 2.66 |
| | | | - | 2.18 | 4.97 |
| | | | | - | 1.91 |
| | | | | | - |

Cuadro 12.4: (a) Parámetros estimados

| proporciones | | | | | |
|---|---|---|---|---|---|
| S1 | S2 | S3 | S4 | S5 | S6 |
| 0.277 | 0.214 | 0.289 | 0.223 | 0..268 | 0.198 |
| OR | | | | | |
| - | 7.23 | 9.1 | 5.42 | 5.53 | 6.90 |
| | - | 5.74 | 5.95 | 6.53 | 7.87 |
| | | - | 5.90 | 3.93 | 7.66 |
| | | | - | 5.40 | 11.56 |
| | | | | - | 5.86 |
| | | | | | - |

Cuadro 12.5: (b) Reparametrización a proporciones y OR

Los intervalos de confianza para los parámetros que surgen del modelo (intensidades y asociaciones) se calculan utilizando la normalidad asintótica de los estimadores máximo verosímiles con la siguiente formulación:

$$[\phi - Z_{(1-\alpha/2)} * s.e; \phi + Z_{(1-\alpha/2)} * s.e] \tag{12.20}$$

donde $s.e.$ es la raíz cuadrada de la varianza de cada parámetro del modelo, la que se estima para cada caso, a través de la descomposición QR de la hessiana asociada al modelo.

| intervalos de confianza al 95 % para las intensidades | | | |
|---|---|---|---|
| intensidades | Ext. Inf. | Estimación puntual | Ext. Sup. |
| 1 | 0.065 | 0.097 | 0.129 |
| 2 | 0.041 | 0.066 | 0.091 |
| 3 | 0.086 | 0.124 | 0.161 |
| 4 | 0.048 | 0.076 | 0.103 |
| 5 | 0.091 | 0.130 | 0.169 |
| 6 | 0.026 | 0.045 | 0.065 |

Cuadro 12.6: intervalos de confianza al 95 % para las intensidades

| intervalos de confianza al 95 % para las asociaciones | | | |
|---|---|---|---|
| asociaciones | Ext. Inf. | Estimación puntual | Ext. Sup. |
| 1-2 | 1.284 | 2.678 | 4.072 |
| 1-3 | 2.517 | 4.675 | 6.833 |
| 1-4 | 0.720 | 1.585 | 2.450 |
| 1-5 | 1.247 | 2.423 | 3.598 |
| 1-6 | 0.802 | 1.871 | 2.939 |
| 2-3 | 0.886 | 1.880 | 2.874 |
| 2-4 | 0.794 | 1.777 | 2.761 |
| 2-5 | 1.447 | 2.859 | 4.270 |
| 2-6 | 1.095 | 2.513 | 3.931 |
| 3-4 | 1.017 | 2.140 | 3.262 |
| 3-5 | 0.634 | 1.271 | 1.908 |
| 3-6 | 1.202 | 2.665 | 4.128 |
| 4-5 | 1.071 | 2.182 | 3.293 |
| 4-6 | 2.358 | 4.970 | 7.581 |
| 5-6 | 0.871 | 1.916 | 2.961 |

Cuadro 12.7: intervalos de confianza al 95 % para las asociaciones

### 12.3.3. Discusión

Puede verse en este caso que según el modelo ajustado, el sextante con mayor intensidad (parcial) es el $S5$ con un valor de $\hat{\phi}_5 = 0.13$ mientras que los sextantes que presentan mayor asociación (parcial) son el $S4, S6$ y el $S1, S3$ que son los sextantes posteriores inferiores y superiores respectivamente, con valores de $\hat{\alpha}_{4,6}$ =4.97

y $\hat{\alpha}_{1,3}$ =4.67.

Si se opta por reducir el número de parámetros del modelo mediante restricciones de igualdad, surgen diferentes alternativas. Una posibilidad es el modelo de "independencia", en dicho caso se impone $\alpha_{ij} = 1 \ \forall \ i, j$, logrando así que solo se estimen las $k$ intensidades del modelo. Otro caso donde se simplifica la dimensionalidad del modelo es el caso de "homogeneidad", en este caso se asume $\alpha_{ij} = \alpha_{kl}$ de modo que se estimen $k$ intensidades y una sola asociación, común a todos los pares de sextantes. Utilizando una prueba de cociente de verosimilitud, se pudieron contrastar las hipótesis de estos modelos. A continuación se presentan las líneas de código:

```
# para testear la hipotesis de asociaciones=1 (independencia)
modelo1.indep<-estim(x,restr=c(rep(NA,6),rep(1,15)))
1-pchisq(-2*(modelo1.indep$Logv-modelo1$Logv),df=p*(p-1)/2)
# para testear la hipotesis de asociaciones iguales (homogeneidad?)
modelo1.homog<-estim(x,restr=c(rep(NA,6),rep(-1,15)))
1-pchisq(-2*(modelo1.homog$Logv-modelo1$Logv),df=p*(p-1)/2-1)
```

Para el caso de la independencia entre los sextantes, se pudo rechazar la independencia ya que el valor del estadístico (cuya distribución era $\chi^2_{15}$) arrojó un valor $p \leq$0.00. Para el caso del modelo de homogeneidad de asociaciones, el estadístico de prueba tiene un grado de libertad menos debido a que se estima un parámetro de asociación. En este caso el $p$-valor fue de 0.043, rechazando así, que todas las asociaciones fuesen iguales a un único valor desconodico. Por lo tanto en ambos casos se rechazan la independencia y la homogeneidad de asociaciones.

En última instancia, retomando que en el modelo sin restricciones se observó que las estimaciones de las asociaciones posteriores (sextantes S1-S3 y sextantes S4-S6) eran mucho mayores al resto, se decidió poner a prueba la siguiente hipótesis:

$$\alpha_{13} = \alpha_{46}$$

Para esto, se ajustó un nuevo modelo bajo esta restricción. Al comparar las verosimilitudes, el p-valor encontrado fue de 0.942, lo que sugirió que las asociaciones posteriores eran efectivamente, de la misma magnitud.

```
\# para testear la hipotesis alfa13 = alfa46
restriccion<-rep(NA,15)
restriccion[c(2,14)]<- -1
restriccion<-c(rep(NA,6),restriccion)
modelo1.restr<-estim(x,restr=restriccion)
modelo1.indep<-estim(x,restr=c(rep(NA,6),rep(1,15)))
1-pchisq(-2*(modelo1.indep\$Logv-modelo1\$Logv),df=p*(p-1)/2)
```

## 12.4. Conclusiones y futuros pasos

En este trabajo se presenta una metodología de análisis para varias variables binarias diferente a la que habitualmente se usa y que está basada en una descomposición de una distribución Bernouilli Multivariada en términos que reflejan

*intensidades* de cada variable y *asociaciones* entre estas, que ya se había presentado por primera vez con resultados también en forma preliminar sobre enfermedad periodontal en el documento de trabajo [2] .

Para el caso de una aplicación en salud oral se analizan las asociaciones entre sextantes en el sangrado.

1. Se descartó la hipótesis de que la presencia de sangrado es independiente entre algunos sextantes.

2. Se constató que la asociación de presencia de sangrado entre los sextantes posteriores no difiere entre mandíbula y maxilar.

A futuro se intentará establecer diferentes tipologías que den cuenta del gradiente de infección usando diferentes técnicas a ser combinadas con la distribución Bernoulli Multivariada .

1. Creación de tipologías de sangrado gengival a través de variables latentes que indican la pertenencia a diferentes grupos usando el algoritmo (EM).

2. Clustering a partir de particiones difusas mediante medidas de entropía: [1],[10],[13]

Por otra parte, resta estudiar cómo hacer el proceso de ajuste de los modelos al trabajar con valores faltantes. Esto problema de datos faltates es frecuente en la evaluación de la enfermedad periodontal, cuando existen sextantes que no pueden ser evaluados por no tener las personas las piezas que componen cada sextante.

Resta a su vez poder implementar el cálculo de los intervalos de confianza para las reparametrizaciones de los componentes del modelo (odds, y OR), en donde la varianza debe ser estimada mediante simulación Monte Carlo.