Ciencias Matemáticas, Vol. 33, No. 2, Pag. 105-117, 2019
Recibido 02-2019, Aceptado 02-2019

Tablas de Contingencia y Modelos Log-Lineales. Una Aplicación en un Problema de Salud. Contingency Tables and Log-Linear Models. An Aplication in a Health Problem.

Brenda Lambert Lamazares[1], Vivian Sistachs Vega[2]

Resumen Generalmente suele obviarse el trabajo con variables cualitativas a pesar de que actualmente son estas las que predominan en la mayoría de las investigaciones. En el presente trabajo se estudian dos de las herramientas estadísticas más usadas en el análisis de datos categóricos: las tablas de contingencia y los modelos log-lineales. Se expone también una aplicación de estos métodos la cual consiste en el análisis de diferentes síntomas y características en pacientes que presentan la enfermedad de Síndrome de Guillain-Barré. Mediante el análisis anterior es posible detectar asociaciones entre las variables lo que facilita el diagnóstico de dicha enfermedad y contribuye al estudio de la misma.

Abstract In general the work with qualitative variables is usually ignored although these are currently the variables that predominate in most of the research. In this work are studied two of the statistical tools most used in the analysis of categorical data: the contingency tables and the log-linear models. An application of these methods is also exposed which consists in the analysis of different symptoms and characteristics in patients who have the disease of Guillain-Barré's Syndrome. Thanks to the previous analysis, it is possible to detect associations between the variables, which facilitates the diagnosis of said disease and contributes to its study.

Palabras Clave
Variables Categóricas — Tablas de Contingencia — Modelos Log-Lineales — Síndrome de Guillain-Barré

[1] *Departamento de Matemática, Instituto de Cibernética Matemática y Física , La Habana, Cuba, brenda@icimaf.cu*
[2] *Departamento de Matemática Aplicada, Universidad de La Habana, La Habana, Cuba, vivian@matcom.uh.cu*

Introducción

Diariamente el ser humano categoriza todo lo que le rodea, afecta o incluye, desde aspectos tan generales como la religión o el partido político al que pertenecen un cierto grupo de personas, hasta cuestiones tan particulares como la ropa que usa o la comida que ingiere. Todas estas propiedades, las cuales suelen adoptar diferentes valores, reciben el nombre de variables categóricas.

Para estudiar las relaciones que pueden existir entre las variables categóricas suelen usarse las tablas de contingencia multidimensionales acompañadas del cálculo de diferentes medidas. No obstante, el estudio en tablas de contingencia presenta algunas limitaciones, entre las que se encuentran:

- Incapacidad para probar los casos de independencia ya sea conjunta o parcial entre las variables categóricas en el experimento.
- Incapacidad para realizar el análisis simultáneo de la asociación de pares de variables.
- Desconocer la posibilidad de interacciones de tercer orden y de orden más alto entre las variables.

Ante estas limitaciones se plantean los modelos log-lineales como una alternativa de análisis. El estudio de estos modelos avanzó considerablemente en los últimos cuarenta años del siglo pasado [3].

Entre los estudios estadísticos que se realizan sobre variables categóricas empleando principalmente las tablas de contingencia y los modelos log-lineales se destacan aquellos motivados por necesidades en el área de la salud y la medicina.

Los planteamientos anteriores constituyen la motivación de este trabajo, el cual tiene como objetivo principal el estudio de los métodos y procedimientos estadísticos que pueden ser aplicados a datos categóricos, principalmente de aquellos que estén involucrados tanto con las tablas de contingencia como con la formulación de modelos log-lineales. Es de interés también utilizar los métodos estadísticos anteriores para realizar un estudio retrospectivo sobre pacientes que presentaron la enfermedad del *Síndrome de Guillain-Barré* con el fin de buscar posibles asociaciones entre diferentes síntomas y características presentes en los pacientes. Lo anterior permitirá ofrecer pautas útiles para realizar el diagnóstico de dicha enfermedad

y avanzar en su estudio.

El trabajo se encuentra dividido en tres secciones principales. Las dos primeras secciones estarán dedicadas al estudio de la teoría fundamental que involucra el tratamiento con tablas de contingencia y la formulación e interpretación de los modelos log-lineales respectivamente. En la tercera sección se mostrará una aplicación de la teoría estudiada en un problema de salud.

1. Tablas de Contingencia

Cuando se trabaja con variables categóricas los datos suelen organizarse en tablas de doble entrada, en las que cada entrada representa un criterio de clasificación. Como resultado de esta clasificación, las frecuencias aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas se les denomina tablas de contingencia.

En estadística, las tablas de contingencia, se emplean para registrar y analizar la relación entre dos o más variables, habitualmente de naturaleza cualitativa, ya sean nominales u ordinales.

A no ser que se especifique lo contrario, en lo siguiente se hará referencia a tablas de contingencia de dos dimensiones por las bondades que estas ofrecen para analizar diferentes conceptos.

1.1 Definiciones y Conceptos Fundamentales

Sea una muestra compuesta por N individuos sobre los que se pretende analizar simultáneamente dos atributos. Se designa por $A_1,...,A_p$ y por $B_1,...,B_q$ las p y q modalidades de los atributos A y B respectivamente y por n_{ij}, donde $i = 1,...,p$ y $j = 1,...,q$, la cantidad de individuos que presentan a la vez las modalidades A_i y B_j. La tabla de contingencia que describe a estos N individuos será una tabla de doble entrada como la siguiente:

$A \setminus B$	B_1	...	B_j	...	B_q	n_{i+}
A_1	n_{11}	...	n_{1j}	...	n_{1q}	n_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	n_{i1}	...	n_{ij}	...	n_{iq}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_p	n_{p1}	...	n_{pj}	...	n_{pq}	n_{p+}
n_{+j}	n_{+1}	...	n_{+j}	...	n_{+q}	n_{++}

Se designa por n_{i+} y por n_{+j} los totales marginales de la muestra por fila y por columnas respectivamente, además, el tamaño total de la muestra se denotará indistintamente por N o n_{++}, puesto que $N = n_{++} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$

Se debe tener en cuenta que cada individuo presenta solo una modalidad de cada atributo.

Sea $\pi_{ij} = P(A_i, B_j) = n_{ij}/N$ la probabilidad de que un elemento aleatoriamente escogido pertenezca a la celda ubicada en la fila i y columna j. Al conjunto de todas las probabilidades $\{\pi_{ij}\}$ se le denomina **distribución conjunta**. Estas satisfacen que $\sum_{i=1}^{p}\sum_{j=1}^{q} \pi_{ij} = 1$.

Las **distribuciones marginales** son los totales por filas, $\{n_{i+}\}$, y los totales por columnas, $\{n_{+j}\}$, las cuales se obtienen sumando las proporciones n_{ij}; o sea $n_{i+} = \sum_{j=1}^{q} n_{ij}$ y $n_{+j} = \sum_{i=1}^{p} n_{ij}$. Se interpreta a n_{i+} como el número de veces que aparece la modalidad i-ésima de A con independencia de cual sea la modalidad de B y análogamente para n_{+j}. Las probabilidades marginales son entonces $p_{i+} = n_{i+}/N$ y $p_{+j} = n_{+j}/N$. Estas satisfacen que $\sum_{i=1}^{p} p_{i+} = \sum_{j=1}^{q} p_{+j} = 1$.

A partir de una tabla de contingencia es posible formar también otro tipo de distribuciones denominadas **distribuciones condicionadas**, debido a que para su obtención es preciso definir previamente una condición. Esta condición hará referencia a la fijación *a priori* de una o varias modalidades de una de las variables cualitativas, para posteriormente calcular la distribución de la otra variable cualitativa sujeta a esa condición.

Las tablas de contingencia pueden seguir tres tipos diferentes de esquemas muestrales dependiendo de los elementos que se mantengan fijos o se dejen variar aleatoriamente. En lo que sigue se hará referencia a tablas de contingencia cuyo esquema muestral es multinomial debido a que en el problema que se analizará el tamaño de la muestra tomada es fijo y las variables analizadas toman más de dos valores [10].

1.2 Independencia y Asociación de variables cualitativas

El grado de relación existente entre dos variables no puede ser establecido simplemente observando las frecuencias de una tabla de contingencia por lo que existen varios **estadísticos**, también llamados **medidas descriptivas**, los cuales exponen en forma estandarizada y de fácil comprensión las características del conjunto de datos que se desea describir, permitiendo así una comparación cabal de dos o más grupos de datos.

En el caso de variables cualitativas la falta de independencia suele denominarse **asociación** y el análisis del grado de asociación entre las variables tiene fuerte incidencia en la estadística de atributos.

Se dice que dos atributos son **independientes** cuando entre ellos no existe ningún tipo de influencia mutua. Si dos atributos, A y B, son independientes estadísticamente, la frecuencia relativa conjunta será igual al producto de las frecuencias marginales respectivas. Para que A y B sean independientes habrá de cumplirse que $n_{ij} = (n_{i+}n_{+j})/N$ para todo i, j. En la práctica basta con que la relación se verifique para

$(p-1)(q-1)$ valores de n_{ij}, ya que entonces se verificará para todos los restantes.

Si se designa por m_{ij} la frecuencia esperada que correspondería en el caso de que ambos atributos fuesen independientes, o sea, $m_{ij} = (n_{i+}n_{+j})/N$, se define el ***Coeficiente de Contingencia Chi-Cuadrado de Pearson***, o simplemente ***Chi-Cuadrado***, como sigue:

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(n_{ij}-m_{ij})^2}{m_{ij}} \tag{1}$$

El *Coeficiente de Contingencia Chi-Cuadrado de Pearson* se utiliza para realizar un contraste formal para la hipótesis nula de independencia de los atributos A y B frente a la hipótesis alternativa que presupone la existencia de asociación entre dichos atributos. El contraste se basa en que, bajo la hipótesis nula, el estadístico *Chi-Cuadrado* se distribuye según una χ^2 con $(p-1)(q-1)$ grados de libertad ([2], [10]).

Para realizar el contraste se halla el valor k tal que, siendo α el nivel de significación, $P(\chi^2_{(p-1)(q-1)} \geq k) = \alpha$. Si el valor del estadístico *Chi-Cuadrado* para los datos dados de la tabla de contingencia es mayor que k se rechaza la hipótesis nula de independencia de los atributos A y B al nivel fijado α. En caso contrario se acepta la independencia [5].

Cuando el tamaño de la muestra es menor que 40 o cuando los individuos que muestran una combinación de variables posibles son menores que 5, se calcula en lugar del estadístico *Chi-Cuadrado de Pearson* el ***Estadístico Exacto de Fisher*** [10]. Este ofrece, basándose en la distribución hipergeométrica y en la hipótesis de independencia, la probabilidad exacta de obtener las frecuencia de hecho obtenidas o cualquier otra combinación más alejada de la hipótesis de independencia [2]. En el caso en que las dos variables observadas sean independientes la probabilidad F de obtener cualquier disposición de las n_{ij} según el *Estadístico Exacto de Fisher* viene dada por:

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

Además de los estadísticos anteriores, existe también el estadístico ***Razón de Verosimilitud*** usualmente utilizado para estudiar la relación entre variables categóricas en los modelos log-lineales ([2], [3]). Se calcula de la forma siguiente:

$$RV = 2\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} log\left(\frac{n_{ij}}{m_{ij}}\right) \tag{2}$$

Este también se distribuye según una χ^2 y se interpreta de igual forma que el estadístico *Chi-Cuadrado*.

Como concepto contrario al de independencia se tiene el de asociación. Se dice que A y B están **asociados** cuando aparecen juntos en mayor número de casos que el que cabría esperar si fuesen independientes. Según que esa tendencia a coincidir o no esté más o menos marcada se tendrán distintos grados de asociación. Para medirlos se han ideado diversos procedimientos denominados **coeficientes de asociación**.

Un coeficiente de asociación que se encuentra en la literatura concerniente a este tema es el coeficiente ***Phi*** [13], también llamado ***Coeficiente de Asociación de Mathews***. Este se obtiene de las siguiente forma:

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

donde χ^2 se refiere al valor del estadistico *Chi-Cuadrado*.

En las tablas con dos variables dicotómicas ϕ toma valores entre 0 y 1, pero en tablas con variables con más de dos categorías este coeficiente puede tomar valores mayores que 1 ya que χ^2 puede ser mayor que el tamaño muestral. Su valor es 0 cuando existe una carencia absoluta de asociación entre las variables, o sea, cuando son independientes; y su valor se aproxima a 1 cuando estas muestran una total asociación entre sí.

Otra medida de relación estadística es el ***Coeficiente V de Cramer*** [13], este se calcula incluyéndole una ligera modificación a la medida anterior:

$$V = \sqrt{\frac{\chi^2}{min(p-1,q-1)\,N}}$$

Este constituye una medida simétrica para la intensidad de la relación entre dos o más variables de la escala nominal cuando, al menos, una de las dos variables tiene, como mínimo, dos valores posibles. Su valor es independiente del tamaño de la muestra y siempre se encontrará entre 0 y 1, donde se obtendrá como resultado el valor 1 cuando exista una relación perfecta entre las variables y el valor 0 cuando estas sean independientes. Para un valor de este estadístico mayor o igual que 0.3 se dirá que existe una correlación significativa y si este es mayor que 0.6 esta correlación será clasificada como relativamente intensa.

Existen también una serie de medidas de asociación utilizadas en los casos en que las variables categóricas sean ordinales. Estas utilizan la información ordinal que las medidas diseñadas para datos nominales pasan por alto [8].

Cuando se trabaja con datos ordinales tiene sentido hablar de la dirección de la relación: una **dirección positiva** indica que los valores altos de una variable se asocian con los valores altos de la otra variable y los valores bajos con los valores bajos; mientras que una **dirección negativa** indica que los valores altos de una variable se asocian con los valores bajos de la otra y viceversa.

Muchas de las medidas de asociación diseñadas para estudiar la relación entre variables ordinales se basan en el concepto de inversión y no inversión [1]. Si los dos valores de un caso en ambas variables son mayores (menores) que los dos valores de otro caso, se dice que entre esos casos se da una **no inversión**. Si el valor de un caso de las variables es mayor que el de otro caso, y el valor del segundo caso es mayor que el del primero, se dice que se da una **inversión**. Finalmente, se plantea que se da un **empate** si dos casos tienen valores idénticos en una o en las dos variables.

Nótese que cuando predominan las no inversiones, la relación es positiva pues conforme aumentan (disminuyen) los valores de una de las variables, aumentan (disminuyen) los valores de la otra. Procediendo análogamente se puede inferir que la relación es negativa cuando predominan las inversiones.

La medida ***Gamma de Goodman y Kruskall*** se basa en la relación relativa que siguen los rangos de dos variables categóricas expresados en escala ordinal, es decir, hace referencia a la inversión o no entre los rangos de los atributos para los individuos observados [13]. Su valor viene definido por la expresión:

$$\gamma = \frac{n_P - n_Q}{n_P + n_Q}$$

donde n_P es el número de no inversiones, es decir, pares de observaciones en que los rangos de ambos factores siguen la misma dirección y n_Q es el número de inversiones, es decir, pares de observaciones en que los rangos de ambos factores siguen direcciones opuestas.

Los valores de γ oscilan ente -1 y 1, alcanzando los extremos cuando se tiene la perfecta asociación negativa o positiva respectivamente. Si dos variables son independientes se obtiene $\gamma = 0$ y si no lo son, la asociación será tanto mayor cuanto más se aproxime γ, en valor absoluto, a la unidad.

Por último se estudiarán los ***Coeficientes de Correlación por Rangos de Kendall*** τ_b y τ_c. Estas medidas se calculan de acuerdo a las fórmulas siguientes:

$$\tau_b = \frac{n_P - n_Q}{\sqrt{(n_P + n_Q + n_{E(A)})(n_P + n_Q + n_{E(B)})}} \quad \tau_c = \frac{2t(n_P - n_Q)}{N^2(t-1)}$$

donde n_P y n_Q representan los valores antes señalados, $n_{E(A)}$ y $n_{E(B)}$ hacen referencia al número de pares empatados en la variable A o B respectivamente, t se refiere al mínimo entre el número de filas y el número de columnas y N será el tamaño de la muestra.

Los valores de ambos coeficientes se encuentran entre -1 y 1, pero τ_b alcanza los valores extremos solo en las tablas de contingencia cuadradas donde ninguna frecuencia marginal vale 0 mientras que τ_c toma valores en este intervalo sin importar el número de filas y de columnas de la tabla. La interpretación de ambos estadísticos se realiza de forma idéntica al estadístico γ discutido anteriormente.

[1] También conocidos como discordancia y concordancia respectivamente.

1.3 Análisis de los Residuos

Otro método existente para detectar las fuentes de asociación entre variables categóricas es el análisis de los residuos en la tabla de contingencia. Se pueden distinguir tres tipos diferentes de residuos: los residuos no tipificados, los tipificados y los tipificados corregidos [8].

Los ***residuos no tipificados*** son las diferencias existentes entre las frecuencias observadas (n_{ij}) y las frecuencias esperadas (m_{ij}) de cada casilla, o sea:

$$r_{ij} = n_{ij} - m_{ij}$$

Se llaman ***residuos tipificados*** al resultado de dividir el residuo no tipificado entre la raíz cuadrada de su correspondiente frecuencia esperada, o sea:

$$\overline{r_{ij}} = \frac{r_{ij}}{\sqrt{m_{ij}}} = \frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}}}$$

Su valor esperado vale 0, pero su desviación típica es menor que 1, lo que hace que no puedan interpretarse como puntuaciones Z. Sin embargo, sirven como indicadores del grado en que cada casilla contribuye al valor del estadístico *Chi-Cuadrado*; de hecho, sumando los cuadrados de los valores de los residuos tipificados se obtiene el valor de dicho estadístico.

Por último, se conoce como ***residuos tipificados corregidos*** al cociente entre el residuo tipificado de cada casilla y su error típico, o sea:

$$\widehat{r_{ij}} = \frac{\overline{r_{ij}}}{\sqrt{V(\overline{r_{ij}})}} = \frac{(n_{ij} - m_{ij})/\sqrt{m_{ij}}}{\sqrt{\left(1 - \frac{n_{i+}}{N}\right)\left(1 - \frac{n_{+j}}{N}\right)}} \tag{3}$$

La gran utilidad de estos residuos radica en que, si asumimos una distribución multinomial y con N suficientemente grande, cada $\widehat{r_{ij}}$ es aproximadamente distribuido como una variable normal estándar por lo que son fácilmente interpretables: utilizando un nivel de confianza del 95%, se puede afirmar que los residuos mayores que 1.96 indican casillas con más casos de los que debería haber si las variables estudiadas fueran independientes; mientras que los residuos menores de -1.96 corresponden a casillas con menos casos de los que cabría esperar bajo la condición de independencia [14].

En tablas de contingencia con variables nominales, una vez que se ha establecido que entre dos variables existe una asociación significativa y se ha cuantificado dicha asociación, los residuos tipificados corregidos constituyen una de las mejores herramientas disponible para poder interpretar con precisión el significado de la asociación detectada.

2. Modelos Log-Lineales

Los problemas que resuelven las tablas de contingencia, siendo importantes, dejan sin respuesta cuestiones de gran importancia como son:

- La estimación de la influencia individual que cada factor ejerce sobre las frecuencias, a través de sus diferentes niveles.
- La cuantificación de la influencia correspondiente a la acción conjunta de varios factores sobre la magnitud de las frecuencias de las celdas en el conjunto de la tabla, en el caso que la contrastación de la independencia no haya conducido a rechazar la hipótesis.

Estos temas son tratados en los modelos log-lineales.

2.1 Modelo Log-Lineal para dos factores

Se comenzará el análisis teórico de los modelos logarítmicos lineales por aquellos que tienen como fin explicar la estructura de tablas de dos dimensiones, debido a que estos son mucho más simples [5]. Para el estudio de modelos que incluyan un número mayor de factores se realiza un análisis análogo.

Los modelos logarítmicos lineales se fundamentan en el principio de independencia de las variables aleatorias que lo componen. Para el caso de una tabla bidimensional, la condición de independencia se puede traducir en que la probabilidad de que un elemento presente simultáneamente cierta características de A y de B en los niveles i-ésimo y j-ésimo, es igual al producto de la probabilidad de que presente la característica de A en el nivel i-ésimo y la de B en el nivel j-ésimo. Basándose en lo anterior se puede entonces explicar el comportamiento conjunto de los dos factores a través de los respectivos comportamientos individuales, o sea:

$$\pi_{ij} = \pi_{i+}\ \pi_{+j}$$

Por el contrario, considerando la no existencia de independencia entre los factores, se puede expresar para todo i y para todo j la probabilidad conjunta como:

$$\pi_{ij} = \pi_{i+}\ \pi_{+j}\ k_{ij} \qquad k_{ij} > 0$$

donde k_{ij} constituye la cuantificación del efecto conjunto del nivel i-ésimo del factor A y del j-ésimo del factor B. A este efecto conjunto se le dará el nombre de **interacción de ambas variables** o **factores**.

En términos de las frecuencias esperadas, la independencia puede expresarse como:

$$m_{ij} = N\ \pi_{i+}\ \pi_{+j} \tag{4}$$

y la no independencia:

$$m_{ij} = N\ \pi_{i+}\ \pi_{+j}\ k_{ij} \qquad k_{ij} > 0 \tag{5}$$

Si se toman logaritmos en las expresiones anteriores, se llega a que, en el caso de independencia

$$\log\ m_{ij} = \log\ N + \log\ \pi_{i+} + \log\ \pi_{+j} \tag{6}$$

y cuando la independencia no existe

$$\log\ m_{ij} = \log\ N + \log\ \pi_{i+} + \log\ \pi_{+j} + \log\ k_{ij} \tag{7}$$

Las expresiones 6 y 7 constituyen una primera formulación de los modelos logarítmico lineales para el caso de dos factores.

Aunque los modelos multiplicativos iniciales (4 y 5) son, respectivamente, equivalentes a los modelos logarítmicos aditivos (6 y 7) , resulta más razonable utilizar los últimos pues en estos el término que discrimina la existencia o no de independencia es $\log\ k_{ij}$, el cual toma el valor cero cuando se está en presencia de independencia, valor más acorde con la idea de ausencia de asociación que se pretende representar, en vez del valor unidad que se le asignaba al parámetro k_{ij} en el modelo multiplicativo inicial para este mismo caso [14].

2.1.1 Efectos Principales e Interacciones

Teniendo en cuenta que la frecuencia esperada para la casilla (i, j) es una función aditiva de un efecto de fila $i-$ésimo y de un efecto de columna $j-$ésimo y efectuando una serie de transformaciones elementales con el objetivo de hacer más operativo el modelo general, se obtiene una nueva formulación del modelo log-lineal de dos factores:

$$\log\ m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

donde:

$$\lambda = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\log\ m_{ij}$$

$$\lambda_i^A = \frac{1}{q}\sum_{j=1}^{q}\log\ m_{ij} - \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\log\ m_{ij}$$

$$\lambda_j^B = \frac{1}{p}\sum_{i=1}^{p}\log\ m_{ij} - \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\log\ m_{ij}$$

$$\lambda_{ij}^{AB} = \log\ m_{ij} - \frac{1}{p}\sum_{i=1}^{p}\log\ m_{ij} - \frac{1}{q}\sum_{j=1}^{q}\log\ m_{ij} + \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\log\ m_{ij}$$

El sumando λ representa la media general de los logaritmos de las $p\ q$ frecuencias estimadas. Este cuantifica el valor que adoptarían los logaritmos de dichas frecuencias si los dos factores, A y B, no ejercieran ningún efecto. El sumando λ_i^A mide la influencia que la fila i-ésima ejerce sobre el logaritmo del número de elementos que posean ese nivel de A. De manera análoga, el sumando λ_j^B evalúa el efecto que el nivel

j-ésimo del factor B ejerce sobre la aparición de elementos en ese nivel.

Los parámetros λ_i^A y λ_j^B reciben el nombre de **efectos principales** o **directos**. Los valores positivos de los efectos directos indican que el nivel en cuestión actúa favoreciendo la presencia de individuos en esa fila o columna; o sea, que la aparición de individuos en esa posición es alta. Como es de esperar, los valores negativos indican la situación contraria, o sea, el nivel no favorece, sino penaliza, la presencia de individuos en ese nivel.

Nótese que los parámetros λ_i^A y λ_j^B son las diferencias entre la media de la distribución marginal de los logaritmos de las frecuencias esperadas, respecto al factor B o A respectivamente, y la media general λ. Estas diferencias dicen que los efectos producidos por los distintos niveles del factor A o B se miden a través de las desviaciones de las medidas marginales de cada nivel respecto a la media general, eliminando así el efecto general que existiría si todas las celdas tuvieran el mismo número de elementos.

Si se suman los p términos de λ_i^A y los q términos de λ_j^B se obtiene que ambas sumas son iguales a 0. Lo anterior tiene como consecuencia que los efectos producidos por los niveles de una característica, tal y como han sido definidos, no son independientes entre sí.

Los parámetros λ_{ij}^{AB} reciben el nombre de **interacciones** y expresan la cuantificación de la acción conjunta del nivel i-ésimo del factor A y del nivel j-ésimo del factor B.

Al igual que los efectos principales, las interacciones de los niveles de cada factor verifican ciertas restricciones. En el caso particular que se está tratando las restricciones son las siguientes:

$$\sum_{i=1}^{p} \lambda_{ij}^{AB} = 0 \qquad \sum_{j=1}^{q} \lambda_{ij}^{AB} = 0$$

Cada una de las dos sumas marginales son iguales a 0, lo que supone la independencia de los efectos de las interacciones.

2.2 Independencia y Asociación en Modelos Log-Lineales

Anteriormente se vió la formulación de un modelo log-lineal con dos variables categóricas bajo la hipótesis de independencia y de no independencia. Este modelo es fácilmente generalizable para cualquier número de variables. Debido a la aplicación que se expondrá resulta de interés presentar el modelo log-lineal para cuatro variables y estudiar la independencia o asociación que pueden existir entre sus componentes [11], por lo que a esta tarea se le dedicará esta sección.

Primeramente véanse algunas definiciones importantes:

Se llama *modelo saturado* a aquel que incluye todos los efectos principales e interacciones posibles entre las variables. Este contiene en su formulación tantos parámetros independientes como celdas tenga la tabla de contingencia a la que es aplicado. Nótese que, el modelo saturado, reproduce exactamente las frecuencias observadas. Si no está presente alguno de los parámetros en el modelo, a este se le denotará *modelo no saturado* ([2], [3]).

Se llama *modelo jerárquico* a aquel modelo no saturado que es coherente en el sentido de que, si en el modelo falta un término, en este estarán excluidos también todos los parámetros de orden superior que contengan la combinación de subíndices fijos del que no aparece [7].

Para ilustrar los conceptos de independencia y asociación se hará uso del modelo saturado de cuatro dimensiones:

$$\begin{aligned}\log\, m_{ij} = \lambda &+ \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{il}^{AD} \\ &+ \lambda_{jk}^{BC} + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} + \lambda_{ijk}^{ABC} + \lambda_{ijl}^{ABD} + \lambda_{ikl}^{ACD} \\ &+ \lambda_{jkl}^{BCD} + \lambda_{ijkl}^{ABCD}\end{aligned}$$

Nótese que en este modelo están presentes una serie de efectos que contribuyen a la tendencia de un elemento de pertenecer a una u otra celda de la tabla. Estos efectos son: los efectos principales ($\lambda_i^A,\ \lambda_j^B,\ \lambda_k^C,\ \lambda_l^D$), las interacciones de primer orden ($\lambda_{ij}^{AB},\ \lambda_{ik}^{AC},\ \lambda_{il}^{AD},\ \lambda_{jk}^{BC},\ \lambda_{jl}^{BD},\ \lambda_{kl}^{CD}$), las interacciones de segundo orden ($\lambda_{ijk}^{ABC}, \lambda_{ijl}^{ABD}, \lambda_{ikl}^{ACD}, \lambda_{jkl}^{BCD}$) y por último, la interacción de orden tres, que constituye el efecto conjunto de las cuatro características (λ_{ijkl}^{ABCD}).

Se puede decir que los cuatro factores son **completamente independientes** si lo son tanto simultáneamente, como dos a dos y tres a tres; o sea que, para i, j, k y l, todas las interacciones deberán ser nulas. Esto conduce a que el modelo resultante sea:

$$log\ m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D$$

Recurriendo al modelo de dimensión cuatro se plantea que el factor A es **completamente independiente** de los otros tres factores cuando todas las interacciones en las que aparece este factor son nulas, obteniéndose el siguiente modelo:

$$log\ m_{ij} = \lambda + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{jk}^{BC} + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} + \lambda_{jkl}^{BCD}$$

Partiendo una vez más del modelo saturado, se puede decir que los factores A y B presentan **independencia condicional** cuando son independientes entre sí para cada nivel de los dos factores restantes, pudiendo estar ambos factores asociados a su vez con un tercer factor, teniendo entonces el modelo resultante al sustituir que: $\lambda_{ij}^{AB} = \lambda_{ijk}^{ABC} = \lambda_{ijl}^{ABD} = \lambda_{ijkl}^{ABCD} = 0$

Finalmente, se dirá que el modelo presenta **asociación parcial** entre los cuatro factores cuando las interacciones de orden tres son nulas siendo diferentes de 0 todas las interacciones de orden dos.

2.3 Estimación de los parámetros en un Modelo Log-Lineal

Los estimadores de los parámetros λ en los modelos log-lineales pueden obtenerse, en ocasiones, sustituyendo las frecuencias esperadas estimadas en las ecuaciones que definen dichos parámetros. En el caso en que no puedan ser obtenidas de forma directa se usan diferentes métodos iterativos, de los cuales, expondremos dos a continuación.

2.3.1 Método de Ajuste Iterativo Proporcional

Este método tiene importantes propiedades entre las que se encuentran [10]:

- Siempre converge al único conjunto de estimaciones máximo verosímil.
- Puede usarse una regla de parada que asegure exactitud para cualquier grado deseado.
- De existir estimaciones directas, el procedimiento produce el estimador exacto en solo un ciclo.
- Cualquier conjunto de evaluaciones iniciales puede ser escogido conforme al modelo inicial fijado.

Su procedimiento, para un modelo log-lineal de cuatro factores, puede resumirse en los siguientes pasos:

Paso 1: $\widehat{m}_{ijkl}^{(0)} = 1 \qquad \forall i,j,k,l.$

Paso 2: $\widehat{m}_{ijkl}^{(4r+1)} = \frac{n_{ijk+}}{\widehat{m}_{ijk+}^{(4r)}}\, \widehat{m}_{ijkl}^{(4r)}$

Paso 3: $\widehat{m}_{ijkl}^{(4r+2)} = \frac{n_{ij+l}}{\widehat{m}_{ij+l}^{(4r+1)}}\, \widehat{m}_{ijkl}^{(4r+1)}$

Paso 4: $\widehat{m}_{ijkl}^{(4r+3)} = \frac{n_{i+kl}}{\widehat{m}_{i+kl}^{(4r+2)}}\, \widehat{m}_{ijkl}^{(4r+2)}$

Paso 5: $\widehat{m}_{ijkl}^{(4(r+1))} = \frac{n_{+jkl}}{\widehat{m}_{+jkl}^{(4r+3)}}\, \widehat{m}_{ijkl}^{(4r+3)}$

Este proceso culminará cuando la diferencia máxima entre las frecuencias observadas y esperadas en las tablas marginales correspondientes a los estadísticos suficientes sea menor que algún δ prefijado.

2.3.2 Método de Newton-Raphson

El *Método de Newton-Raphson* es un procedimiento general para encontrar el máximo de una función de varias variables. En el caso de los modelos log-lineales se maximizará la función de verosimilitud tratada como una función de los parámetros del modelo, o sea, $L(\beta)$. Este método nos provee de una secuencia de aproximaciones de los parámetros del modelo las cuales convergen al máximo local de la función siempre que, tanto la función como la aproximación inicial de los parámetros sean lo suficientemente buenas [2].

En más detalle, mostremos como el *Método de Newton-Raphson* halla el valor $\widehat{\beta}$ donde la función $L(\beta)$ se maximiza.

Sea $u' = (\partial L(\beta)/\partial\beta_1, \partial L(\beta)/\partial\beta_2, ...)$ y sea H la matriz hessiana de la verosimilitud, denotaremos por $u^{(t)}$ y $H^{(t)}$ las evaluaciones en $\beta^{(t)}$ del vector u y de la matriz H respectivamente, donde $\beta^{(t)}$ representa la $t-$ésima aproximación de β.

En el paso t del proceso iterativo se aproxima $L(\beta)$ por su desarrollo en Series de Taylor de orden dos, obteniéndose:

$$L(\beta) \approx L\left(\beta^{(t)}\right) + u^{(t)'}\left(\beta - \beta^{(t)}\right) + \frac{1}{2}\left(\beta - \beta^{(t)'}\right) H^{(t)} \left(\beta - \beta^{(t)}\right)$$

y se resuelve $\partial L(\beta)/\partial\beta \approx u^{(t)} + H^{(t)}\left(\beta - \beta^{(t)}\right) = 0$ en función de β para hallar su próxima aproximación. Dicha aproximación se puede expresar como:

$$\beta^{(t+1)} = \beta^{(t)} - \left(H^{(t)}\right)^{-1} u^{(t)}$$

asumiendo que la matriz $H^{(t)}$ sea no singular.

El proceso continua iterando hasta que las variaciones entre $L(\beta^{(t)})$ y $L(\beta^{(t+1)})$ son lo suficientemente pequeñas. El estimador máximo verosímil de los parámetros será $\beta^{(t)}$ cuando $t \to \infty$; sin embargo, si la función tiene otro máximo local en el cual su derivada se anule pueden ser seleccionadas como aproximaciones finales valores erróneos, radica en esto la importancia de una buena aproximación inicial.

Los cálculos con el *Método de Newton-Raphson* son más complejos que con el *Método de Ajuste Iterativo Proporcional* ya que es necesario invertir matrices en cada ciclo pero la convergencia de este es más rápida y el mismo es aplicable a una gama más amplia de problemas, razones por las cuales es el método que aparece implementado en la mayoría de los softwares matemáticos y estadísticos [10].

2.4 Selección de un Modelo Log-Lineal Adecuado

Cuando trabajamos con un problema particular existen varios modelos log-lineales que lo describen, ejemplo de esto son el modelo saturado y los modelos minimales. El modelo saturado, al incluir todos los posibles parámetros, presenta un ajuste perfecto a nuestros datos pero su interpretación suele ser muy compleja. Los modelos minimales, al contener el menor número de parámetros permitidos, suelen tener interpretaciones muy sencillas pero generalmente no presentan un buen ajuste. Surge entonces la problemática de buscar un equilibrio entre la bondad de ajuste del modelo y la dificultad en su interpretación. Para la solución de la problemática planteada anteriormente se han planteado varios métodos entre los que se encuentran el *Método Particionado* y el *Método Paso a Paso*.

2.4.1 Método Particionado

Este método consta de los pasos siguientes [10]:

1- Se construye una jerarquía de modelos.
2- Se particiona el estadístico *Razón de Verosimilitud*, tanto en el modelo general como en los modelos sucesivos, siempre desde el modelo más complejo hasta el más simple.
3- Se selecciona el nivel de significación deseado.
4- Se realizan pruebas de bondad de ajuste a las componentes particionadas, comenzando con el modelo más complejo. Se agrega la componente sobre la que se ha realizado la prueba.
5- Se repite el paso anterior hasta que sea significativa la diferencia entre modelos sucesivos o lo sea la suma de las componentes acumuladas

Este procedimiento suele presentarse en forma de tabla y depende de la jerarquía de modelos seleccionada, en la que gran cantidad de modelos no son considerados, lo que constituye un inconveniente en tablas de altas dimensiones. Otro inconveniente que presenta es que es un procedimiento muy engorroso de realizar manualmente por lo que es aconsejable en su lugar programarlo.

2.4.2 Método Paso a Paso

Este procedimiento fue propuesto en 1971 por Goodman y modificado 6 años más tarde. Consiste en un procedimiento de selección ascendente y resulta de gran utilidad en tablas de grandes dimensiones [10].

Este método consta de los pasos siguientes:

1- Se establece un nivel de significación.
2- Se examina la bondad de ajuste de los modelos jerárquicos comenzando por los modelos que contemplan hasta los efectos de la primera variable, posteriormente hasta los efectos de la segunda variable y así sucesivamente hasta el modelo saturado. En este paso nos detendremos cuando encontremos un modelo que contemple todos los efectos de d variables y que ajuste bajo las condiciones impuestas.
3- Al modelo que contempla los efectos de $d-1$ variables y no ajusta a nuestros datos se le adicionarán y sustraerán parámetros λ al nivel d en dependencia de su contribución individual al ajuste del modelo.
4- Se repite el paso anterior hasta que que no puedan adicionarse ni eliminarse parámetros λ del modelo en cuestión.

2.5 Inferencia en Modelos Log-Lineales

La finalidad principal del análisis de modelos log-lineales es identificar las relaciones más significativas que se producen entre las variables de estudio, generalmente numerosas medidas calculadas con este fin son altamente dependientes del modelo en cuestión, razón por la cual resulta de gran importancia la correcta selección del modelo.

Existen varios estadísticos que cuantifican en que medida se ajusta adecuadamente un modelo concreto a la estructura de una tabla de contingencia dada. Las dos medidas más usuales para esto son el estadístico *Chi-Cuadrado de Pearson* y la *Razón de Verosimilitud* cuyas fórmulas (1 y 2) fueron planteadas anteriormente. Para el caso en el que se analizan cuatro variables, se pueden extender estas fórmulas directamente a las siguientes [5]:

$$\chi^2 = \sum_{ijkl} \frac{(n_{ijkl} - m_{ijkl})^2}{m_{ijkl}} \qquad RV = 2\sum_{ijkl} n_{ijkl} log\left(\frac{n_{ijkl}}{m_{ijkl}}\right)$$

Ambos estadísticos distribuye según una χ^2 donde los grados de libertad están determinados por la diferencia entre el número de celdas de la tabla de contingencia y la cantidad de parámetros del modelo. Los grados de libertad disminuyen a medida que el modelo se vuelve más complejo.

Después de calcular los estadísticos anteriores se realiza una prueba de hipótesis donde:

H_0 : *El modelo seleccionado se ajusta a los datos.*
H_a : *El modelo seleccionado no se ajusta a los datos.*

A pesar de que las pruebas anteriores proporcionan una idea global de cómo los datos de la tabla de contingencia proceden de una población que obedece al modelo log-lineal seleccionado, estas no informan acerca de la relevancia particular en ese ajuste de los parámetros del modelo, ni de los casos extremos que pueden darse en cada una de las celdas de clasificación cruzada. Es por esto que, complementando estas pruebas, se debe realizar un análisis basado en los residuos del modelo y en su grado concreto de significación estadística [7].

Estos residuos se calculan siguiendo la fórmula 3 y se interpretan al igual que se hacia en el análisis en tablas de contingencia. Ellos, en este nuevo contexto, muestran la calidad del ajuste celda a celda. Se debe tener presente que es posible que una celda muestre falta de ajuste en un modelo bien ajustado.

El uso del análisis residual permite ordenar tanto los efectos principales como las interacciones entre las variables de acuerdo con su grado de significación y su magnitud. Este análisis posibilita también reducir el modelo cuando algún parámetro sea significativamente igual a 0, marcando así la pauta del modelo no saturado que parezca más adecuado [11].

3. Aplicación en un Problema de Salud

Como se planteó anteriormente son disimiles las aplicaciones en el área de la salud y la medicina de técnicas estadísticas y probabilísticas. Estas tienen diversos fines, entre ellos se encuentran el estudio de una enfermedad particular o de sus síntomas, el análisis de la reacción de los pacientes ante un

medicamento determinado, el esclarecimiento de ciertos patrones en el progreso de una enfermedad o la detección del tratamiento más apropiado para una dolencia bajo ciertas características; todos ellos tributan enormemente a mejorar la vida de los seres humanos y a la comprensión de numerosos trastornos biológicos que nos afectan. En la presente sección mostraremos como la teoría estudiada puede ser aplicada para encontrar relaciones entre los síntomas de la enfermedad del Síndrome de Guillain-Barré con el objetivo de avanzar en su estudio y facilitar su diagnóstico.

3.1 Síndrome de Guillain-Barré

El Síndrome de Guillain-Barré (SGB) es la neuropatía aguda más frecuente, de evolución más rápida y potencialmente fatal [9]. Su gravedad varía desde una debilidad ligera en los miembros inferiores hasta la cuadriplejía flácida con parálisis respiratoria y trastornos disautonómicos graves que pueden conducir al enfermo a la muerte.

El SGB suele afectar a personas de cualquier edad y sexo con dos picos de presentación: uno en la etapa adulta joven (15-34 años) y otro en ancianos (64-70 años). Este se presenta en pocas ocasiones en niños menores de un año de edad [1]. Este síndrome puede ser difícil de diagnosticar en sus primeras semanas debido a que varios trastornos presentan síntomas similares [12]. La enfermedad evoluciona en tres fases, denominadas: fase de progresión, de estabilización y de regresión (o de recuperación) las cuales suelen completarse en un periodo cuya duración oscila de 3 a 6 meses [4].

La mortalidad estimada del SGB es variable y aún con el advenimiento de una terapia efectiva sigue siendo de un 4 % a un 8 %. Aproximadamente, un 80 % de las personas que presentaron SGB se recuperan adecuadamente después del tratamiento, sin embargo, la calidad de vida de estas personas puede verse dañada en diferentes áreas muchos años después del inicio de la enfermedad, indicando recuperación incompleta a largo plazo en aproximadamente un 10 % de los pacientes [6].

A pesar de que existen tratamientos que reducen la gravedad de los síntomas y aceleran la recuperación en la mayoría de los pacientes, tales como la plasmaféresis y el tratamiento con altas dosis de inmunoglobulinas [15], no hay una cura específica para esta enfermedad. También se desconocen métodos efectivos para su prevención.

Por todo lo anteriormente planteado, resulta beneficioso realizar un estudio con pacientes que presentaron esta enfermedad con el objetivo de analizar diferentes síntomas y características de la misma en aras de detectar una posible asociación entre ellos.

3.2 Diseño Muestral y Metodológico

Se realizó un estudio retrospectivo de la enfermedad del Síndrome de Guillain-Barré en 143 pacientes tratados antes de septiembre del 2017 en el *Instituto de Neurología y Neurocirugía* donde se buscaron relaciones en dos grupos de cuatro variables cada uno.

El primer grupo estuvo conformado por las cuatro variables cualitativas nominales siguientes:

Diag_seg: Se refiere a los métodos que se emplean para poder diagnosticar la enfermedad. Tiene tres posibles valores: NF (No funcional), LCR (Líquido cefalorraquídeo) y Ambos.

SíntResp: Hace alusión a la presencia de trastornos respiratorios en el paciente. Tiene dos valores posibles: Si y No.

Deb_MInf: Hace referencia a la presencia de debilidad en los miembros inferiores del paciente. Tiene dos valores posibles: Si y No.

Deb_MSup: Hace referencia a la presencia de debilidad en los miembros superiores del paciente. Tiene dos valores posibles: Si y No.

Las variables cualitativas ordinales siguientes conformaron el segundo grupo de estudio:

Per_Prog: Se refiere a la duración del período de progresión de la enfermedad en cada paciente. Tiene tres valores posibles: 1 (de 0 a 7 días), 2 (de 8 a 21 días) y 3 (más de 22 días)

Per_Estab: Se refiere a la duración del período de estabilización de la enfermedad en cada paciente. Tiene los mismos tres valores posibles que la variable anterior.

Per_Recup: Se refiere a la duración del período de recuperación de cada paciente. Tiene los mismos tres valores posibles que la primera variable de este grupo.

NESNI: Acrónimo de **Nueva Escala de Severidad para Neuropatías Inflamatorias**. Hace referencia a una escala que mide la intensidad, severidad y extensión de la enfermedad en los pacientes. Tiene tres valores posibles: 1 (Severo), 2 (Moderado) y 3 (Ligero).

3.3 Resultados y Discusión

Se consideró oportuno dividir el estudio realizado en dos análisis independientes los cuales responden a cada uno de los grupos de variables antes mencionados.

3.3.1 Primer Análisis

Realizando una tabla de contingencia que agrupe al primer grupo de variables y a sus categorías se obtiene la siguiente tabla de orden 4 con dimensión 24 [2] (Cuadro 1).

[2] En la tabla no se incluyen las combinaciones de variables que no agrupen a ningún individuo con el objetivo de reducir la tabla y hacer más fácil su comprensión.

Cuadro 1. Tabla de Contingencia. Primer Análisis

Diag_seg	***SíntResp***			***Deb_Sup*** *Si*	*No*
NF	*Si*	***Deb_MInf***	*Si*	*2*	
	No	***Deb_MInf***	*Si*	*10*	*1*
LCR	*Si*	***Deb_MInf***	*Si*	*8*	
	No	***Deb_MInf***	*Si*	*55*	*10*
			No		*1*
Ambos	*Si*	***Deb_MInf***	*Si*	*3*	
	No	***Deb_MInf***	*Si*	*45*	*6*
			No		*1*

Posteriormente se calculan medidas de asociación adecuadas entre todos los pares posibles de variables obteniéndose que el único par en el que se evidenciaba una clara asociación era en el formado por **Deb_MInf** y **Deb_MSup** pues para esta se obtuvo un valor del *Estadístico Exacto de Fisher* igual a 0.017. Se considera importante también la relación que se establece entre las variables **SíntResp** y **Deb_MSup**, pues al calcular el estadístico *Razón de Verosimilitud* para ellas se obtiene 3,932 cuyo p-valor es 0,047 (Cuadro 2).

Cuadro 2. Pruebas de Asociación Parcial

Comb. Variables	*gl*	*Valor*	*Sig.*	*Estadístico*
Diag_seg × ***Deb_MInf***	*2*	*0.254*	*0.881*	*Chi-Cuad.*
Diag_seg × ***Deb_MSup***	*2*	*0.524*	*0.770*	*Chi-Cuad.*
Diag_seg × ***SíntResp***	*2*	*1.756*	*0.416*	*Chi-Cuad.*
Deb_MInf × ***Deb_MSup***	*1*	*0.017*	*——*	*Fisher*
Deb_MInf × ***SíntResp***	*1*	*0.387*	*0.534*	*RV*
Deb_MSup × ***SíntResp***	*1*	*3.932*	*0.047*	*RV*

Se comprueba, basándose en los valores y la significación de las medidas de asociación calculadas, que entre las variables **Deb_MInf** y **Deb_MSup** existe una correlación significativa (Cuadro 3).

Cuadro 3. Medidas de Asociación. **D_MInf** × **D_MSup**

Medidas de Asociación	*Valor*	*Sig. Aprox.*
Phi	*0.304*	*0.000*
V de Cramer	*0.304*	*0.000*

Otro indicio de esta asociación son los valores absolutos mucho mayores de 1.96 de los residuos tipificados corregidos de la tabla de contingencia perteneciente a estas variables (Cuadro 4).

Cuadro 4. Tabla de Contingencia. **D_MInf** × **D_MSup**

			Deb_Sup *Si*	*No*	*Total*
Deb_MInf	*Si*	*Recuento*	*123*	*17*	*140*
		Res. Tip. Corr.	*3.6*	*-3.6*	
	No	*Recuento*	*0*	*2*	*2*
		Res. Tip. Corr.	*-3.6*	*3.6*	
Total		*Recuento*	*123*	*19*	*142*

Centrándose en la búsqueda del modelo log-lineal que se ajuste a los datos, se realiza primeramente la prueba de los K-efectos. Esta, en su doble versión, permite comprobar si los efectos de un determinado orden son iguales a 0. La primera parte de la prueba contrasta la hipótesis de que todos los efectos de un orden particular o superior son iguales a 0; la segunda parte, contrasta la hipótesis de que solamente los efectos de un determinado orden son iguales a 0. Para ambos casos se rechazaran las interacciones de orden K para las que los p-valores asociados al estadístico *Chi-Cuadrado* sean mayores al nivel de significación. Las interacciones rechazadas no son estadísticamente significativas.

Al realizar la prueba de las K-efectos con los datos estudiados se obtuvo que las interacciones de cuatro y de tres variables no son significativas mientras que la de dos variables y los efectos principales sí lo son (Cuadro 5).

Cuadro 5. Prueba de los K-efectos

			Razón de Veros.		*Pearson*	
	K	*gl*	*Chi-Cuad.*	*Sig.*	*Chi-Cuad.*	*Sig.*
Efectos de orden K superiores	*1*	*23*	*434.759*	*0.000*	*764.930*	*0.000*
	2	*18*	*14.875*	*0.671*	*21.047*	*0.277*
	3	*9*	*0.000*	*1.000*	*0.000*	*1.000*
	4	*2*	*0.000*	*1.000*	*0.000*	*1.000*
Efectos de orden K	*1*	*5*	*419.884*	*0.000*	*743.882*	*0.000*
	2	*9*	*14.875*	*0.094*	*21.047*	*0.012*
	3	*7*	*0.000*	*1.000*	*0.000*	*1.000*
	4	*2*	*0.000*	*1.000*	*0.000*	*1.000*

El hecho de que los efectos de determinadas interacciones o de los efectos principales sean distintos de 0 no implica que, en particular, cada uno de ellos lo sea; por lo que es aconsejable realizar la prueba de asociación parcial. En esta se contrasta la hipótesis de que un efecto en particular es nulo y se mantendrán en el modelo aquellos efectos cuyo p-valor sea menor que 0.05 siempre y cuando no se entre en contradicción con el modelo jerárquico. La prueba de asociación parcial comprueba la significación de cada efecto individual, es decir, su contribución al ajuste del modelo. Nótese que se vuelve a obtener la interacción entre las variables **SíntResp** y **Deb_MSup** como un caso límite (Cuadro 6).

Cuadro 6. Asociaciones Parciales

Efecto	*gl*	*Chi-Cuadrado Parcial*	*Sig.*
Diag_seg × ***SíntResp*** × ***Deb_MInf***	*2*	*0.000*	*1.000*
Diag_seg × ***SíntResp*** × ***Deb_MSup***	*2*	*0.000*	*1.000*
Diag_seg × ***Deb_MInf*** × ***Deb_MSup***	*2*	*0.001*	*0.999*
SíntResp × ***Deb_MInf*** × ***Deb_MSup***	*1*	*0.000*	*1.000*
Diag_seg × ***SíntResp***	*2*	*1.789*	*0.409*
Diag_seg × ***Deb_MInf***	*2*	*0.343*	*0.842*
SíntResp × ***Deb_MInf***	*1*	*0.000*	*1.000*
Diag_seg × ***Deb_MSup***	*2*	*0.529*	*0.768*
SíntResp × ***Deb_MSup***	*1*	*3.587*	*0.058*
Deb_MInf × ***Deb_MSup***	*1*	*7.804*	*0.005*
Diag_seg	*2*	*49.048*	*0.000*
SíntResp	*1*	*109.919*	*0.000*
Deb_MInf	*1*	*175.831*	*0.000*
Deb_MSup	*1*	*85.085*	*0.000*

Para seleccionar el modelo más adecuado se emplea el *Método Paso a Paso*, el cual parte del modelo saturado y va eliminando los efectos no significativos empezando por el de mayor orden deteniéndose cuando no pueda eliminarse

ningún efecto más sin perder poder predictivo.

Se llega a la conclusión de que el modelo final se genera por las interacciones de la variable **Deb_MSup** con **SíntResp** y con **Deb_MInf** y por el efecto principal de la variable **Diag_seg**. Para este modelo se analiza la tabla de frecuencias observadas, esperadas y los residuales, así como también las Pruebas de Ajuste (Cuadro 7) llegando a la conclusión de que este es un buen modelo para ajustar los datos.

Cuadro 7. Pruebas de Ajuste

	Valor	*gl*	*Sig.*
Razón de Verosimilitud	*2.708*	*16*	*1.000*
Chi-Cuadrado de Pearson	*2.473*	*16*	*1.000*

Se obtiene así que el modelo es:

$$log\ m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} \qquad (8)$$

donde A se refiere a la variable **Diag_seg**, B a **SíntResp**, C a **Deb_MInf** y D a **Deb_MSup**.

Se puede ver que la variable **Diag_seg** presenta independencia condicional con todas las variables restantes, así como también la presenta el par de variables **SíntResp** y **Deb_MInf**.

Finalmente se calculan las estimaciones de los parámetros del modelo, y su transformación en puntuaciones Z (Cuadro 8). Los valores absolutos de los parámetros indican la intensidad de la asociación. [3] [4]

Cuadro 8. Estimaciones de los Parámetros

Parámetro	*Estimación*	*Z*	*Sig.*
Constante	*-0.255*	*-0.357*	*0.721*
***Deb_MInf**=1 × **Deb_MSup**=1*	*4.007*	*5.616*	*0.000*
***Deb_MInf**=1 × **Deb_MSup**=2*	*2.140*	*2.863*	*0.004*
***Deb_MInf**=2 × **Deb_MSup**=1*	*-16.439*	*-0.011*	*0.991*
***SíntResp**=1 × **Deb_MSup**=1*	*-2.136*	*-7.282*	*0.000*
***SíntResp**=1 × **Deb_MSup**=2*	*-18.714*	*-0.012*	*0.991*
***Diag_seg**=1*	*-1.442*	*-4.677*	*0.000*
***Diag_seg**=2*	*0.297*	*1.667*	*0.096*

Interpretando el modelo log-lineal resultante (8) se puede inferir que los pacientes están caracterizados por una elevada presencia de debilidad en los miembros inferiores, la cual comúnmente se encuentra acompañada por debilidad en los miembros superiores. Respecto a la debilidad en los miembros superiores y la presencia de síntomas de trastornos respiratorios se infiere que es raro que se presenten simultáneamente en el mismo paciente. Finalmente, se puede afirmar que la frecuencia media en el diagnóstico seguro cuando se hace la prueba del líquido cefalorraquídeo (LCR) es superior a las otras categorías.

[3] Para $|Z| > 1{,}96$ el efecto es significativo al 95 % y para $|Z| > 2{,}57$ lo es al 99 %.

[4] En la tabla no se muestran las estimaciones de los parámetros redundantes pues estos no aportan nueva información.

3.3.2 Segundo Análisis

Para realizar el estudio del segundo grupo de variables se realiza primeramente la tabla de contingencia que agrupe a todas las variables de estudio y a sus categorías, obteniéndose así un tabla de orden 4 y dimensión 81 (Cuadro 9) [5].

Cuadro 9. Tabla de Contingencia. Segundo Análisis

					Per_Estab	
NESNI	***Per_Recup***			*0 - 7*	*8 - 22*	*+ 22*
Severo	*0 - 7*	***Per_Prog***	*0 - 7*	*3*		
	8 - 22	***Per_Prog***	*0 - 7*	*2*		
			8 - 22	*2*	*1*	
Moderado	*0 - 7*	***Per_Prog***	*0 - 7*	*12*		
			8 - 22	*4*	*1*	
			+ 22	*3*		
	8 - 22	***Per_Prog***	*0 - 7*	*10*	*1*	
			8 - 22	*13*	*2*	
			+ 22	*2*	*1*	
	+ 22	***Per_Prog***	*0 - 7*	*4*		
			8 - 22	*5*	*2*	
			+ 22	*1*		
Ligero	*0 - 7*	***Per_Prog***	*0 - 7*	*6*		
			8 - 22	*9*		*1*
			+ 22	*2*		*1*
	8 - 22	***Per_Prog***	*0 - 7*	*8*		
			8 - 22	*11*		
			+ 22	*1*		
	+ 22	***Per_Prog***	*0 - 7*	*6*	*2*	
			8 - 22	*19*	*2*	*1*
			+ 22	*2*		

Al analizar los valores de las pruebas basadas en el estadístico *Chi-Cuadrado de Pearson* en todas las tablas de contingencia que se pueden crear al cruzar solo dos de nuestras variables, se obtiene evidencia de asociación entre la variable **Per_Recup** con **NESNI** y con **Per_Prog** (Cuadro 10).

Cuadro 10. Pruebas de Asociación Parcial

Comb. Variables	*gl*	*Valor*	*Sig.*	*Estadístico*
NESNI** × **Per_Prog	*4*	*5.781*	*0.216*	*RV*
NESNI** × **Per_Estab	*4*	*6.484*	*0.166*	*RV*
NESNI** × **Per_Recup	*4*	*10.667*	*0.030*	*RV*
Per_Prog** × **Per_Estab	*4*	*5.284*	*0.259*	*RV*
Per_Prog** × **Per_Recup	*4*	*9.718*	*0.043*	*RV*
Per_Estab** × **Per_Recup	*4*	*8.005*	*0.091*	*RV*

Al calcular las medidas de asociación nominales se obtuvo que en ambas interacciones existía una asociación débil ya que a pesar de resultar estas significativas sus valores eran menores que 2.8 en ambos casos (Cuadros 11 y 12).

Cuadro 11. Medidas de Asociación. **Per_Prog** × **Per_Recup**

	Medidas de Asociación	*Valor*	*Sig. Aprox.*
Nominal	*Phi*	*0.261*	*0.045*
	V de Cramer	*0.184*	*0.045*
Ordinal	*Tau-b de Kendall*	*0.123*	*0.114*
	Tau-c de Kendall	*0.113*	*0.114*
	Gamma	*0.196*	*0.114*

[5] Las casillas correspondientes a ciertas combinaciones de las variables fueron eliminados pues ellas no contenían a ningún individuo

Cuadro 12. Medidas de Asociación. **Per_Recup** × **NESNI**

	Medidas de Asociación	*Valor*	*Sig. Aprox.*
Nominal	*Phi*	*0.270*	*0.034*
	V de Cramer	*0.191*	*0.034*
Ordinal	*Tau-b de Kendall*	*0.158*	*0.038*
	Tau-c de Kendall	*0.145*	*0.038*
	Gamma	*0.251*	*0.038*

Respecto a las medidas ordinales calculadas, solo resultaron significativas aquellas que representan la interacción entre las variables **Per_Recup** y **NESNI**, las cuales tiene valores absolutos distantes de 1, por lo que se ratifica la suposición de que la asociación entre dichas variables es débil (Cuadro 12).

Se reafirmó la existencia de asociación en estas dos combinaciones de variables al analizar en sus tablas correspondientes los valores de los residuos tipificados corregidos ya que existían casillas con valores absolutos mayores que 1.96 (Cuadros 13 y 14).

Cuadro 13. Tabla de Contingencia. **Per_Prog** × **Per_Recup**

			Per_Recup			
			0 - 7	*8 - 22*	*+ 22*	*Total*
Per_Prog	*0 - 7*	*Recuento*	*21*	*21*	*12*	*54*
		Res. Tip. Corr.	*1.9*	*0.2*	*-2.1*	
	8 - 22	*Recuento*	*15*	*29*	*32*	*76*
		Res. Tip. Corr.	*-2.7*	*0.1*	*2.5*	
	+ 22	*Recuento*	*6*	*4*	*3*	*13*
		Res. Tip. Corr.	*1.4*	*-0.5*	*-0.8*	
Total		*Recuento*	*42*	*54*	*47*	*143*

Cuadro 14. Tabla de Contingencia. **Per_Recup** × **NESNI**

			NESNI			
			Sev.	*Mod.*	*Lig.*	*Total*
Per_Recup	*0 - 7*	*Recuento*	*3*	*20*	*19*	*42*
		Res. Tip. Corr.	*-0.2*	*0.8*	*-0.7*	
	8 - 22	*Recuento*	*5*	*29*	*20*	*54*
		Res. Tip. Corr.	*0.5*	*2.1*	*-2.3*	
	+ 22	*Recuento*	*3*	*12*	*32*	*47*
		Res. Tip. Corr.	*-0.4*	*-2.9*	*3.1*	
Total		*Recuento*	*11*	*61*	*71*	*143*

Enfocándose en la búsqueda del modelo log-lineal que se ajuste a los datos de estudio, se realizaron las pruebas de los *K*-efectos (Cuadro 15) y las pruebas de asociación parcial (Cuadro16).

Cuadro 15. Prueba de los *K*-efectos

			Razón de Veros.		*Pearson*	
	K	*gl*	*Chi-Cuad.*	*Sig.*	*Chi-Cuad.*	*Sig.*
Efectos de orden K superiores	*1*	*80*	*369.467*	*0.000*	*557.678*	*0.000*
	2	*72*	*64.405*	*0.726*	*81.245*	*0.213*
	3	*48*	*19.380*	*1.000*	*18.750*	*1.000*
	4	*16*	*0.399*	*1.000*	*0.235*	*1.000*
Efectos de orden K	*1*	*8*	*305.062*	*0.000*	*476.433*	*0.000*
	2	*24*	*45.025*	*0.006*	*62.495*	*0.000*
	3	*32*	*18.981*	*0.967*	*18.515*	*0.973*
	4	*16*	*0.399*	*1.000*	*0.235*	*1.000*

Cuadro 16. Asociaciones Parciales

Efecto	*gl*	*Chi-Cuadrado Parcial*	*Sig.*
Per_Prog × ***Per_Estab*** × ***Per_Recup***	*8*	*3.976*	*0.859*
Per_Prog × ***Per_Estab*** × ***NESNI***	*8*	*3.431*	*0.904*
Per_Prog × ***Per_Recup*** × ***NESNI***	*8*	*6.554*	*0.585*
Per_Estab × ***Per_Recup*** × ***NESNI***	*8*	*4.204*	*0.838*
Per_Prog × ***Per_Estab***	*4*	*4.455*	*0.348*
Per_Prog × ***Per_Recup***	*4*	*7.223*	*0.125*
Per_Estab × ***Per_Recup***	*4*	*8.267*	*0.082*
Per_Prog × ***NESNI***	*4*	*4.264*	*0.371*
Per_Estab × ***NESNI***	*4*	*7.833*	*0.098*
Per_Recup × ***NESNI***	*4*	*10.251*	*0.036*
Per_Prog	*2*	*50.600*	*0.000*
Per_Estab	*2*	*198.533*	*0.000*
Per_Recup	*2*	*1.518*	*0.468*
NESNI	*2*	*54.410*	*0.000*

Basándose en los resultados anteriormente expuestos resulta lógico proponer un modelo log-lineal que incluya las interacciones de orden dos y los efectos principales de cada variable.

Al emplear el *Método Paso a Paso* se llegó a la conclusión de que el modelo final se genera por las interacciones de la variable **Per_Recup** con **Per_Prog** y con **NESNI** y por el efecto principal de la variable **Per_Estab**.

Después de realizar las Pruebas de Ajuste (Cuadro 17) y de analizar la tabla de frecuencias observadas, esperadas y los residuales se puede concluir que el modelo propuesto se ajusta a los datos.

Cuadro 17. Pruebas de Ajuste

	Valor	*gl*	*Sig.*
Razón de Verosimilitud	*43.901*	*64*	*0.974*
Chi-Cuadrado de Pearson	*54.740*	*64*	*0.789*

Finalmente el modelo será:

$$log\ m_{ijkl} = \lambda + \lambda_i^E + \lambda_j^F + \lambda_k^G + \lambda_l^H + \lambda_{kl}^{GH} + \lambda_{ik}^{EG} \qquad (9)$$

donde E se refiere a la variable **Per_Prog**, F a **Per_Estab**, G a **Per_Recup** y H a **NESNI**.

Se puede ver que la variable **Per_Estab** presenta independencia condicional con todas las variables restantes, así como también la presenta el par de variables **Per_Prog** y **NESNI**.

Para concluir se calculan las estimaciones de los parámetros del modelo, y su transformación en puntuaciones *Z* (Cuadro 18).

Interpretando el modelo log-lineal resultante (9) se puede inferir que la enfermedad del Síndrome de Guillain-Barré generalmente suele estabilizarse en menos de 7 días, siendo muy raro que tarde en hacerlo un período mayor a 22 días. Respecto a la relación entre el tiempo de progresión de la enfermedad y de recuperación del enfermo se puede plantear que nunca

Cuadro 18. Estimaciones de los Parámetros

Parámetro	*Estimación*	*Z*	*Sig.*
Constante	*-3.150*	*-3.849*	*0.000*
***Per_Recup**=1 × **NESNI**=1*	*-1.561*	*-1.725*	*0.085*
***Per_Recup**=1 × **NESNI**=2*	*0.336*	*0.458*	*0.647*
***Per_Recup**=1 × **NESNI**=3*	*0.284*	*0.387*	*0.698*
***Per_Recup**=2 × **NESNI**=1*	*-1.707*	*-1.940*	*0.052*
***Per_Recup**=2 × **NESNI**=2*	*0.050*	*0.065*	*0.949*
***Per_Recup**=2 × **NESNI**=3*	*-0.321*	*-0.406*	*0.685*
***Per_Recup**=3 × **NESNI**=1*	*-2.367*	*-3.920*	*0.000*
***Per_Recup**=3 × **NESNI**=2*	*-0.981*	*-2.898*	*0.004*
***Per_Prog**=1 × **Per_Recup**=1*	*1.253*	*2.706*	*0.007*
***Per_Prog**=1 × **Per_Recup**=2*	*1.658*	*3.040*	*0.002*
***Per_Prog**=1 × **Per_Recup**=3*	*1.386*	*2.148*	*0.032*
***Per_Prog**=2 × **Per_Recup**=1*	*0.916*	*1.897*	*0.058*
***Per_Prog**=2 × **Per_Recup**=2*	*1.981*	*3.714*	*0.000*
***Per_Prog**=2 × **Per_Recup**=3*	*2.367*	*3.920*	*0.000*
***Per_Estab**=1*	*3.746*	*6.412*	*0.000*
***Per_Estab**=2*	*1.466*	*2.289*	*0.022*

será menor el tiempo de recuperación del paciente comparado con el tiempo de progresión de la enfermedad, ya que mientras más lento progrese esta, más tiempo necesitará el paciente para recuperarse. Finalmente, se puede afirmar que los pacientes pertenecientes a las categorías de "Severo" o "Moderado" de la variable **NESNI** suelen tener un período de recuperación cuya duración no excede los 22 días.

Conclusiones

Después de realizada la investigación presentada se puede concluir que el trabajo con tablas de contingencia complementado con el planteamiento de un modelo log-lineal apropiado constituye un método estadístico eficaz para el tratamiento de variables cualitativas y la búsqueda de asociación entre las mismas. Lo anterior propicia que dichos métodos sean utilizados en estudios pertenecientes a áreas no relacionadas directamente con las matemáticas como lo es el área de la salud.

Agradecimientos

A la Dra. Zurina Lestayo OFarril del Instituto de Neurología, por facilitar la información de los datos.

Referencias

[1] Acosta, M; Cañizá , M; Romano, M. y Mateo, E. *Síndrome de Guillain-Barré* Revista de Posgrado de la *VIa* Cátedra de Medicina, 168, 2007.

[2] Agresti A. *An Introduction to Categorical data Analysis.* USA: Jhon Wiley & Sons, 2007.

[3] Agresti A. *Categorical Data Analysis.* USA: Jhon Wiley & Sons, 2013.

[4] *Boletín Informativo sobre el Síndrome de Guillain-Barré.* Organización Mundial de la Salud: Oficina Regional para las Américas & Organización Panamericana de la Salud, 2016.

[5] Christensen, R. *Log-linear models.* USA: Springer, 1990.

[6] *Diagnóstico y Tratamiento del Síndrome de Guillain Barré en el Segundo y Tercer nivel de Atención.* Ciudad de México: Instituto Mexicano del Seguro Social, 2016. Recuperado el 16 de noviembre de 2017 de http://imss.gob.mx/profesionales-salud/gpc

[7] Dobson, A. *An introduction to generalized linear models.* USA: Chapman & Hall/CRC. pp 156-172, 2002.

[8] Fuente, S de la. *Análisis de Variables Categóricas: Tablas de Contingencia.* Facultad de Ciencias Económicas y Empresariales. Universidad Autónoma de Madrid, Notas de Curso. 2011.

[9] Lestayo, Z. y Hernandez, JL. *Análisis del comportamiento del síndrome de Guillain-Barré. Consensos y discrepancias.* Revista Neurología, 46(4), 230-237, 2008.

[10] Linares, G. *Análisis de datos multivariados.* México: Editorial Benemérita Universidad Autónoma de Puebla. pp 197-232, 2016.

[11] McCullagh, P. y Nelder, J. *Generalized Linear Models.* London: Chapman & Hall. pp 193-234, 1989.

[12] Mendoza-Hernández, D; Blancas, L. y Gutiérrez, J. *Síndrome de Guillain-Barré.* Revista Alergia, Asma e Inmunologías Pediátricas, 19(2), 56-63, 2010.

[13] Navarro, R. *Introducción a la bioestadística. Análisis de variables binarias* México: McGraw-Hill, 1988.

[14] Pérez C. *Técnicas de Análisis Multivariante de Datos. Aplicaciones con SPSS.* Madrid: Pearson Education S.A, 2004.

[15] Pérez, J. *Síndrome de Guillain-Barré. Actualización.* Revista Acta Neurología Colombia, 22(2), 201-208, 2006.