

## ¿Cómo asociar variables cualitativas? Un procedimiento ordenado para utilizar pruebas Chi cuadrado de independencia y Regresión Logística

## How to associate qualitative variables? An orderly procedure to use of Chi square tests of independence and Logistic Regression

Arnaldo Mangeaud<sup>1,2</sup> , Haydée Cugno<sup>2</sup>, María Gabriela Molina<sup>2</sup>, David Elías Panigo<sup>2,3</sup>, Natalia Andrea Lozano<sup>4,5</sup>, Mariana Sacco Ramello<sup>4</sup>, Analía Gonzalez<sup>1,2,6</sup>.

1. Universidad Católica de Córdoba, Facultad de Ciencias de la Salud. Unidad de Bioestadística y Metodología

2. Universidad Nacional de Córdoba. Facultad de Ciencias Exactas, Físicas y Naturales. Cátedras de Bioestadística I y II

3. Universidad Católica de Córdoba, Facultad de Ciencias de la Salud. Cátedra Metodología de la Investigación y Bioestadística. Licenciatura en instrumentación quirúrgica.

4. Universidad Católica de Córdoba, Facultad de Ciencias de la Salud. Servicio de Alergia e Inmunología

5. Universidad Católica de Córdoba. Facultad de Ciencias de la Salud. Cátedra Metodología de la Investigación y Bioestadística. Carrera de Medicina.

6. Universidad Católica de Córdoba. Facultad de Ciencias de la Salud. Cátedra de Metodología de la Investigación y Bioestadística. Licenciatura en kinesioterapia y fisioterapia.

Correspondencia: Arnaldo Mangeaud. Email: amangeaud@yahoo.com.ar

### Resumen

La realización de pruebas de Independencia Chi cuadrado o de Regresiones logísticas, a los fines de comprobar hipótesis en los trabajos de investigación, requieren de una serie de pasos desde la detección del rol de cada variable hasta la interpretación de los resultados. El Objetivo del presente es recomendar un procedimiento de pasos ordenados para la realización e interpretación de los análisis estadísticos de variables respuestas de tipo cualitativas, desarrollados particularmente con un ejemplo pediátrico. A partir de la información de historias clínicas de pacientes pediátricos, menores de 2 años, que fueron internados con diagnóstico de bronquiolitis, se realizaron los análisis Chi cuadrado, las regresiones logísticas (simples y múltiples) y la obtención de Odds Ratios. Resultaron significativos los efectos sobre la Asistencia respiratoria mecánica de Score de Tal, Virus Respiratorio Sincitial y frecuencia cardíaca. Se concluye con un manual de procedimientos funcional de 8 pasos a tener en cuenta a los fines realizar un óptimo procedimiento de análisis.

**Palabras claves:** Bioestadística, métodos cuantitativos, test asociaciones, predicciones.

### Abstract

In order to test hypotheses in research work, carrying out Chi-square Independence tests or Logistic Regressions, it requires a series of steps from detecting the role of each variable to interpreting the results. The objective of this document is to recommend a procedure of ordered steps for carrying out and interpreting statistical analyzes of qualitative response variables, developed particularly with a pediatric example. Based on the information from the medical records of pediatric patients, under 2 years of age, who were hospitalized with a diagnosis of bronchiolitis, Chi square analyses, logistic regressions (simple and multiple) and Odds Ratios were obtained. The effects on mechanical respiratory support of Tal Score,

Respiratory Syncytial Virus and heart rate were significant. It concludes with a functional procedures manual of 8 steps to take into account in order to carry out an optimal analysis procedure.

**Keywords:** Biostatistics. quantitative methods, test associations, predictions.

## Introducción

Los trabajos de investigación analíticos, ya sean prospectivos, retrospectivos, observacionales o bien ensayos clínicos, requieren la utilización de herramientas estadísticas específicas que relacionen las variables objeto del estudio con aquellas que los investigadores proponen asociar<sup>1</sup>. Cuando estas variables dependientes o variables “respuestas” son de tipo cualitativas (también llamadas categóricas o nominales), surgen dos tipos de análisis como los más frecuentemente utilizados: la prueba de independencia Chi cuadrado y el análisis de Regresión Logística<sup>2</sup>.

Es extensa la bibliografía que hace referencia a cómo se realizan estas dos pruebas. Si se desea comprender la lógica, el desarrollo y la aplicación de ambos análisis, se puede recurrir a libros generales de estadística<sup>1,3,4</sup>, libros de biostatística<sup>2,5,6</sup>, o bien artículos específicos en literatura de las ciencias de la salud<sup>7,8,9</sup>. En el prolífico mundo de la literatura científica, hasta encontramos discusiones sobre si es correcto denominar Chi cuadrado o Ji cuadrado<sup>10</sup> al análisis desarrollado por Karl Pearson<sup>11</sup>.

Por lo anteriormente expuesto, reiterar dicha información en este artículo sería poco novedoso, por lo que sólo resumiremos los siguientes puntos, que creemos que son importantes a tener en cuenta:

La prueba de independencia Chi cuadrado, se realiza con el objeto de asociar dos variables cualitativas. Es muy común observar los casos donde cada una de ellas posee únicamente dos categorías. De ese modo, se realiza una tabla de doble entrada (tabla de contingencia) conocida como tabla de 2 x 2. La prueba se utiliza también cuando una o ambas variables poseen 3 o más categorías. Pero, cuando esto ocurre, a mayor cantidad de categorías, mayor es la cantidad del número de pacientes (n) necesarios, ya que esta prueba, tiene restricciones para “n” bajos<sup>6</sup>. Si se cuenta con escaso número de pacientes, o están distribuidos de modo que algunas celdas poseen pocos de ellos, debemos utilizar otra forma de cálculo del valor p denominada de un modo general, método de Montecarlo<sup>12</sup>. Al realizar estos análisis se deben obrar con cuidado, ya que algunos softwares estadísticos lo advierten

(RMedic, SPSS), pero otros brindan un resultado incorrecto sin advertencia previa (R, InfoStat). Otra limitación que posee la relación de dos variables con 3 o más categorías en cada una es que, generalmente, se asocian a esta prueba otros estimadores o medidas como Odds Ratio o Riesgo Relativo y éstos sólo pueden calcularse en tablas de 2 x 2<sup>13</sup>.

Por su parte, la Regresión Logística es un análisis donde se pretende que una variable independiente cuantitativa (X) explique la variabilidad de la variable dependiente cualitativa (Y)<sup>14</sup>. Esta última, llamada variable respuesta, debe tener sólo dos categorías. La Regresión Logística simple (con sólo una variable X) ajusta una ecuación y traza una función de estimación o predicción de Y, condicionada a diferentes valores de X.

En la literatura citada previamente, el lector podrá entender cabalmente los análisis, pero, hemos observado que, en la práctica, a los usuarios les resulta complejo el procedimiento del análisis, los pasos secuenciales y la interpretación de los mismos. El detectar deficiencias en el procedimiento de cómo investigadores en formación analizan las variables, nos motivó a plantear como objetivo del presente, recomendar un procedimiento de pasos ordenados para la realización e interpretación de los análisis estadísticos de variables respuestas de tipo cualitativas, desarrollados particularmente con un ejemplo pediátrico.

## Materiales y métodos

Origen de las variables analizadas.

La base de datos fue confeccionada con la información de las historias clínicas de pacientes pediátricos, menores de 2 años, que fueron internados con diagnóstico de bronquiolitis aguda en el Servicio de Pediatría de la Clínica Universitaria Reina Fabiola (Córdoba, Argentina) en los años 2017, 2018 y 2019.

Al momento de la internación, se evaluó en cada paciente: la saturación de Oxígeno (% de saturación de oxígeno obtenido con oxímetro de pulso, con una FiO<sub>2</sub> de 21%), la frecuencia

cardíaca (latidos por minuto) y el puntaje de score de Tal modificado (que es un score que se obtiene mediante la evaluación de la frecuencia respiratoria, la presencia de sibilancias, de cianosis y de tiraje)<sup>15-18</sup>. Para detección de antígenos virales de virus respiratorio sincitial (VRS) y adenovirus, se registró el resultado del análisis de secreciones nasofaríngeas realizado durante las primeras 24hs de internación, mediante técnica de inmunofluorescencia indirecta (IFI). Además, se registró si el paciente requirió Asistencia Respiratoria Mecánica (ARM) durante la internación.

El diseño de estudio fue de cohorte, retrospectivo, transversal y analítico.

Cabe destacar que, dado los fines didácticos del presente artículo, se realizó una selección parcial de variables, que permiten mostrar la secuencia de análisis de las pruebas estadísticas propuestas, y se dejaron de lado las demás variables estudiadas en pacientes internados con bronquiolitis aguda.

Por lo tanto, en el ejemplo analizado, se tiene por meta detectar qué variables o factores son los que permiten predecir si el paciente requerirá ARM.

### VARIABLES DE ESTUDIO

Variable respuesta:

ARM (si/no): variable que registra si el paciente requirió Asistencia respiratoria mecánica durante la internación.

Variables independientes:

Cuantitativas:

Score de Tal (medido como el valor del Score, de 0 a 12)

Saturación de Oxígeno (en %)

Frecuencia Cardíaca al ingreso (en latidos por minuto)

Cualitativas:

Score de Tal8 (si/no-medido como el Score de Tal  $> 8$  o  $\leq 8$ )

VRS (si/no)

Adenovirus (si/no)

### ANÁLISIS ESTADÍSTICO

La base de datos se elaboró en un archivo Excel. En la primera fila se colocaron los identificadores de pacientes y el nombre de cada una de las variables. Cada columna correspondió a cada una de las variables y cada fila (a partir de la fila 2) correspondió a cada paciente.

Se utilizó en todos los análisis como variable “respuesta” a la variable Asistencia respiratoria mecánica (ARM) y se probaron las asociaciones

entre ésta y el resto de las variables independientes. Secuencialmente se realizaron los siguientes análisis:

Se elaboraron tablas descriptivas.

A los fines de asociar ARM con variables cualitativas, se realizaron Pruebas de Independencia Chi cuadrado (asintótica o por método de Montecarlo, según correspondiera). Además, dado el diseño retrospectivo empleado, se obtuvieron los Odds Ratio.

A los fines de asociar ARM con variables cuantitativas, se realizaron Regresiones logísticas simples y se obtuvieron los Odds Ratio.

A los fines didácticos de comprender la similitud de resultados entre ambas pruebas se realizaron con ARM y las variables categóricas regresiones logísticas simples.

A los fines de detectar cuáles son las mejores variables predictivas se realizó una Regresión Logística múltiple con método stepwise (ingreso secuencial de variables significativas en el modelo)<sup>19, 20</sup>.

La totalidad de los análisis fueron realizados con el software RMedic<sup>21</sup>, con excepción de Regresiones logísticas múltiples (stepwise) que se realizaron con el software SPSS<sup>22</sup>. En todos los casos, el nivel de significación utilizado fue del 5%.

### ASPECTOS ÉTICOS

El estudio se realizó cumpliendo con la normativa de la Declaración de Helsinki, Buenas Prácticas Clínicas de la Administración Nacional de Medicamentos, Alimentos y Tecnología (ANMAT) y la Ley de la provincia de Córdoba N.º 9694 sobre el sistema de evaluación, registro y fiscalización de las investigaciones en salud. Se asegura la protección de datos personales de los pacientes según la Ley nacional de la República Argentina N.º 25326. Los autores declaran no tener conflictos de intereses. El estudio ha sido aprobado por la Secretaría de Investigación y Vinculación Tecnológica de la Facultad de Ciencias de la Salud de la Universidad Católica.

### RESULTADOS

#### 1. Variables:

A partir de los objetivos del proyecto surgen variables a analizar. Ya sea que la información se tome in situ en los pacientes o se realice una búsqueda en las historias clínicas institucionales, cada una de las variables que se utilizarán debe estar medida del mismo modo, o expresada con

la misma métrica. Dicho de otro modo, si la variable es un número, toda la variable estará representada por números, si la variable es nominal, serán sólo letras. Tres ejemplos:

Ejemplo 1: la edad del paciente. Debe expresarse en la misma unidad, es decir 6, 4, 8. Siempre significan lo mismo: años. Además, en la base de datos, no se coloca la unidad de medida, es decir se registra un “2”, no se registra “2 años”. Tampoco debe ocurrir que en algunos casos la edad este registrada en años y en otros meses. Es común observar en algunas bases de datos el siguiente error: un 2 que corresponde a dos años y un 3 que corresponde a tres meses, es un error grave.

Ejemplo 2: Edad puede estar expresada en años (2, 4, 6) o expresada también en categorías (< 1 año, entre 1 y 2 años, entre 3 y 6 años), pero siempre de la misma forma. Se debe prestar atención al valor de corte, por ejemplo, es un error no aclarar en qué categoría se incluirá determinada edad: menor a 1 año y mayor a 1 año: ¿dónde quedan los niños de exactamente 1 año?

Ejemplo 3: Presencia o ausencia se puede entender como Si/No o como 1/0, donde 1 es presencia y 0 es ausencia. Siempre se lo debe escribir de la misma manera.

En nuestro caso las variables de estudio fueron tomadas: ARM, como 1 o 0; Frecuencia Cardíaca, como latidos por minuto; Saturación de Oxígeno, como %; Score de Tal, como puntaje de 0 a 12; Categorías de Score de Tal, como 1 a pacientes con puntaje mayor a 8 (bronquiolitis aguda grave) y como 0 a aquellos con valores menores o iguales a 8 (bronquiolitis aguda leve a moderada).

## 2. Base de datos:

Ordenadamente se completa una tabla con la información de cada paciente en filas y las variables en columnas. La primera fila es la única que posee nombre de las variables (Figura 1). Y cada paciente sólo ocupa una fila. Debe destacarse que, si una variable no es conocida para algún paciente, esa celda quedará como un espacio vacío.

	A	B	C	D	E	F	G	H
1	ID	ARM	ScoreTAL	SaturacionO2	FCIngreso	ScoreTAL8	VRS	Adenovirus
2	41	1	7	90	158	1	0	1
3	79	0	8	88	180	0	1	1
4	80	0	6	95	140	1	1	1
5	87	1	6	87	150	1	1	0
6	105	0	8	88	180	0	1	1
7	107	1	9	96	140	0	1	1
8	117	1	7	85	148	1	1	0
9	149	0	8	90	180	0	1	1
10	45	0	3	92	120	1	0	0

**Figura 1.** Corresponde a la base de datos a analizar, pacientes en filas y variables en columnas.

## 3. Variable respuesta:

Se debe identificar cuál es la variable “respuesta” y cuáles son las variables independientes. Si hubiera más de una variable respuesta se debe hacer un listado de variables independientes para cada variable respuesta, ya que no necesariamente en nuestros objetivos deseamos probar que todas las variables independientes pueden influir en todas las variables respuestas. En nuestro caso poseemos sólo una variable respuesta: ARM y poseemos 5 variables independientes: FC, Saturación de O2, Score de Tal, VRS y Adenovirus. Pero como decidimos también contemplar el Score de Tal como mayor o igual a 8, entonces eso hace que sean 6 las variables independientes a analizar.

## 4. Frecuencias absolutas y relativas:

Para entender la frecuencia (o prevalencia) de cada una de variables, siempre es importante realizar una tabla simple con frecuencias absolutas (número de pacientes) y frecuencias relativas (porcentajes o proporciones) de cada una de ellas. Luego, en el informe final que realicemos, intentaremos resumir esa información pasando de múltiples tablas simples a sólo una tabla más compleja. En nuestro caso omitiremos estas tablas a los fines de no distraer la lectura del objetivo de este artículo.

## 5. Prueba Chi cuadrado:

A los fines de revisar cuál o cuáles de las variables cualitativas se relaciona con nuestra variable respuesta realizamos ordenadamente una prueba chi cuadrado para cada una de ellas. Nuestros resultados fueron:

### 5a. Score de Tal (expresada como $\leq 8$ o $> 8$ )

La asociación entre el Score de Tal (TAL8) y la Asistencia respiratoria mecánica (ARM), puesta a prueba con el test Chi cuadrado de independencia, resultó significativa ( $p < 0.001$ ). El test nos dice: hay una asociación entre TAL8 y ARM. Esto nos lleva a la necesidad de observar e interpretar los porcentajes de casos bajo cierta configuración de la tabla para “ver” cómo es esa asociación. La (Tabla 1) presenta la forma de dicha asociación. Por consenso colocamos en filas a “la causa”, la variable predictora o variable independiente, es decir TAL8, y en columna colocamos “la consecuencia” o la variable respuesta, o sea ARM. Como se planteaba anteriormente, es conveniente colocar los porcentajes condicionados por filas (que sumen un 100% por filas). La (Tabla 1) nos presenta la unión de los tres análisis realizados por separado.

**Tabla 1.** Frecuencias absolutas y relativas de Score de Tal (>8), VRS y Adenovirus, con respecto al requerimiento de ARM y sus correspondientes Odds Ratio (Intervalo de confianza 95%) y valor p para la prueba Chi cuadrado.

		ARM				Total	OR (IC95%)	Valor p
		NO		SI				
		paciente s	%	pacientes	%			
TAL8	≤ 8	152	92,68	12	7,32	164	7,24	<0.0001
	> 8	14	63,64	8	36,36	22	(2,45-20,66)	
VRS	No	97	94,17	6	5,83	103	3,19	0.0313
	Si	76	83,52	15	16,48	91	(1,18-8,61)	
Adenovirus	No	169	89,89	19	10,11	188	4,45	0.1369
	Si	4	66,67	2	33,33	6	(0,76-25,93)	

Así, observamos lo siguiente: mientras que en los pacientes con TAL8 ≤ 8, el 7.32% requirió ARM, en cambio, en los pacientes con TAL8 > 8, lo hicieron el 36.36% y eso está respaldado por una asociación significativa.

El Odds ratio de esta asociación arrojó un valor de 7.24 y un intervalo de confianza del 95% de 2.45 a 20.66. La interpretación es que: los pacientes que se internan por bronquiolitis aguda con un puntaje de score de TAL > 8 al momento del ingreso hospitalario tienen 7.24 (IC95% 2.45 – 20.66) veces más de posibilidades de requerir ARM que los pacientes con puntaje de score de TAL < 8.

Es importante comprender que ahora que observamos una asociación significativa, la prevalencia de ARM debe estar condicionada a TAL8, ya que la prevalencia de ARM con Tal > 8, es del 36.36% (es decir, 8 de 22 pacientes requirieron ARM), mientras que la prevalencia de ARM en la población total, sin tener en cuenta TAL8, es de 7.32%, es decir que 12 de 164 pacientes requirieron ARM durante la internación (los pacientes que requirieron ARM dividido el total de pacientes).

El cálculo de Odds Ratio también puede ser interpretado de otra manera, para la cual es importante tener en cuenta lo siguiente: Se conoce que el Odds Ratio con resultado igual a uno indica que ambas categorías de la variable tienen la misma posibilidad de ocurrencia en función de la otra variable, por lo que se deduce que si tengo un Odds Ratio con resultado igual a 2 (dos) indica que cierta categoría de la variable tiene 100% más de posibilidades de ocurrir que la otra categoría.

Dada esta afirmación conceptual, se puede decir que tener un score de TAL > 8 al momento del ingreso hospitalario tiene 624% más de posibilidades de requerir ARM que tener un score ≤ 8.

### 5b.VRS

La asociación entre VRS (Si/No) y el requerimiento de ARM, puesta a prueba con el test Chi cuadrado de independencia, resultó significativa (p=0.0313). La Tabla 1 muestra la forma de dicha asociación. Así, observamos que, mientras que los pacientes sin VRS, el 5.83% requirieron ARM, en cambio, en los pacientes con presencia de VRS, lo hicieron el 16.48% y eso está respaldado por una asociación significativa. El Odds ratio de esta asociación arrojó un valor de 3.19 y un intervalo de confianza del 95% de 1.18 a 8.61. La interpretación es que los pacientes que se internan por bronquiolitis por VRS tienen 3.19 (IC95% 1.18 – 8.61) veces más de posibilidades de requerir ARM que los pacientes con bronquiolitis con VRS negativo. De otra manera se puede interpretar, afirmando que los pacientes con bronquiolitis por VRS tienen 219% más de posibilidades de requerir ARM que los pacientes con bronquiolitis con VRS negativo.

### 5c. Adenovirus (expresada como Si/No)

La asociación entre la presencia de Adenovirus y el requerimiento de Asistencia respiratoria mecánica (ARM), puesta a prueba con el test Chi cuadrado de independencia, resultó no significativa (p=0.119). De todos modos, en la (Tabla 1), observamos que, mientras que los pacientes sin Adenovirus, el 10.11% requirió ARM, en los pacientes con Adenovirus, lo hicieron el 33.33%, pero esta asociación no es significativa. El Odds ratio de esta asociación no significativa arrojó un valor de 4.45 y un intervalo de confianza del 95% de 0.76 a 25.91. Dado que el intervalo de confianza incluye el valor 1, la interpretación es que no hay diferencia en las posibilidades de requerir ARM entre los pacientes que se internan por bronquiolitis con adenovirus positivo en comparación con aquellos que tienen adenovirus negativo (OR= 4.45, IC95% 0.76 – 25.91).

### 6. Regresión Logística simple:

A los fines de constatar cuáles de las variables cuantitativas se relacionan con nuestra variable respuesta, realizamos ordenadamente una Regresión Logística simple para cada una de ellas. Nuestros resultados fueron:

#### 6a. Score de Tal (expresada en un score de 0 a 12)

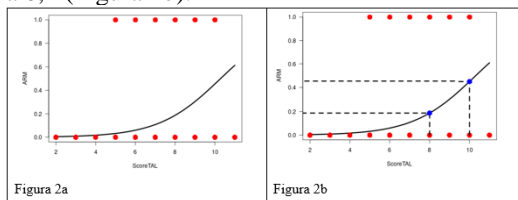
Se observa una dependencia de la ARM con Score de Tal, puesta a prueba con la Regresión

Logística simple ( $p < 0.001$ ). La (Tabla 2) muestra los estimadores de la regresión.

Las variables cualitativas serán las variables denominadas “X” y la variable respuesta se debe codificar en números (Si=1 y No=0) y se denominará “Y”. El valor de la pendiente de cada modelo nos brinda información sobre la relación existente entre cada variable cuantitativa evaluada en nuestro estudio (Score de Tal, Frecuencia cardíaca y saturación de O<sub>2</sub>) y la variable respuesta cualitativa (requirió ARM Si/No). Si el valor de esta pendiente es positivo, quiere decir que cuando X se incrementa, hay más probabilidad de ocurrencia del 1 (la ocurrencia del evento estudiado, en este caso, ARM), en caso contrario, (si la pendiente es negativa) disminuye la probabilidad de la ocurrencia del evento. El valor de la ordenada en el modelo de Regresión Logística casi nunca posee una interpretación sencilla, ya que sería un indicador de la probabilidad del evento (en este caso, requerir ARM) cuando X es igual a cero, lo cual suele ser biológicamente imposible.

Con el valor de la pendiente, matemáticamente, se puede obtener el Odds ratio de esta relación. Para este caso el OR estimado es: 1.9. Este se obtiene con la siguiente fórmula: Odds Ratio =  $\exp(0.6418) = 1.9$ .

Por su parte, la (Figura 2a) representa la dispersión de los datos y la ecuación ajustada. Partimos sabiendo que la pendiente de esta función es significativa, es decir cuando se incrementa Score de Tal, se incrementa ARM, pero como ARM es un Si/No (o bien un 1/0), la interpretación aplicada a este caso es la siguiente: cuando se incrementa Score de Tal, se incrementa la probabilidad de requerir ARM. Dicho esto, se observa en el gráfico que cuando el paciente posee un Score de Tal de 8, la probabilidad de requerir ARM es menor a 0,2, mientras que, si el paciente posee un score de Tal de 10, la probabilidad de requerir ARM es mayor a 0,4 (Figura 2b).



**Figura 2a, 2b.** Relación entre el valor del Score de Tal y ARM mediante Regresión Logística. 2a. Curva ajustada ( $p < 0,001$ ). 2b. Estimación de valores de ARM en función de dos valores Score de Tal (8 y 10).

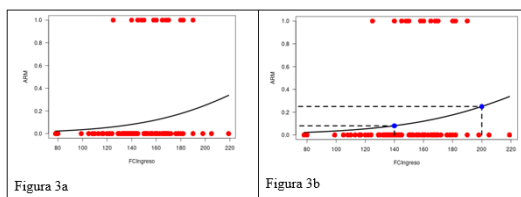
¿Cómo debemos hacer para estimar las probabilidades de ARM condicionadas a un valor del Score de Tal? La respuesta es: debemos realizar un cálculo matemático un tanto complicado. La versión de RMedic actualmente en uso (<https://rmedic.ucc.edu.ar/>) permite realizar la estimación del valor esperado. De hecho, el valor estimado para un Score de Tal de 8 es una probabilidad de 0,187 y el estimado para un Score de Tal de 10 es de 0.454.

#### 6b. Frecuencia cardíaca al ingreso

Se observa una dependencia de la ARM con Frecuencia cardíaca al ingreso, puesta a prueba con la Regresión Logística simple ( $p = 0.040$ ). La (Tabla 3) muestra los estimadores de la regresión. A su vez, con el valor de la pendiente, se puede obtener el Odds ratio de esta relación, para este caso el valor es: 1.02. Por su parte, la figura 3a representa la dispersión de los datos y la ecuación ajustada. Partimos sabiendo que la pendiente de esta función es significativa, es decir cuando se incrementa frecuencia cardíaca al ingreso, se incrementa ARM, decimos que: cuando se incrementa la frecuencia cardíaca al ingreso, se incrementa la probabilidad de requerir ARM. Dicho esto, se observa que cuando el paciente posee una frecuencia cardíaca de 140, la probabilidad de requerir ARM no llega a 0,08, mientras que, si el paciente posee una frecuencia cardíaca de 200, la probabilidad de requerir ARM es de 0,25 (Figura 3b).

#### 6c. Saturación de O<sub>2</sub>

No se observa una dependencia de la ARM con saturación de O<sub>2</sub> al ingreso, puesta a prueba con la Regresión Logística simple ( $p = 0.179$ ). La (Tabla 2) muestra los estimadores de la regresión. A su vez, con el valor de la pendiente, se puede obtener el Odds ratio de esta relación, para este caso el valor es: 0.93. Por su parte, la Figura 4 representa la dispersión de los datos y la ecuación ajustada. Partimos sabiendo que la pendiente de esta función no es significativa, es decir cuando se incrementa saturación al ingreso, no cambia significativamente ARM, por lo tanto, no debemos interpretar nada más. La pendiente no es significativamente distinta de cero y El Odds Ratio no difiere de 1 (Uno).



**Figura 3a y b.** Relación entre el valor del Frecuencia Cardíaca y ARM mediante Regresión Logística. 3a. Curva ajustada ( $p=0.040$ ). 3b estimación de valores de ARM en función de dos valores de Frecuencia Cardíaca (140 y 200).

**Tabla 2.** Resultados de los modelos de Regresión Logística simple entre Score de Tal, Frecuencia Cardíaca y saturación O<sub>2</sub>, con respecto al requerimiento de ARM y su correspondiente valor p para las pruebas.

		Estimador	Error estándar	Valor Wald	OR (IC 95%)	Valor p
Score de Tal	Ordenada	-6.6023	1.3036	-5.06	1.90	< 0.001
	Pendiente	0.6418	1.1682	3.82	(1.36-2.64)	< 0.001
Frecuencia cardíaca	Ordenada	-5.5345	1.7254	-3.21	1.02	< 0.001
	Pendiente	0.0222	0.0108	2.05	(1.001-1.044)	0.040
Saturación O <sub>2</sub>	Ordenada	5.0083	5.2748	0.95	0.93	0.342
	Pendiente	-0.0775	0.0577	-1.34	(0.83-1.04)	0.179

**Tabla 3.** Resumen de los resultados obtenidos en las relaciones mediante variables cualitativas (Chi cuadrado) y cuantitativas (Regresión Logística simple).

Variables cualitativas	Valor p	Odds ratio	Variables cuantitativas	Valor p	Odds ratio
Score de Tal (\$)	<0,001	7,24	Score de Tal (score de 1 a 12)	<0,001	1,90
VRS	0,0313	3,19	FC	0,0402	1,02
Adenovirus	0,119	4,45	Saturación	0,179	0,93

#### 6d. Resumen de resultados parciales y nuevas preguntas

Ya sabemos que presentaron una asociación significativa con ARM las siguientes variables: Score de Tal (en sus dos formas de medirlo), VRS y FC. Y también sabemos que tanto Adenovirus como Saturación no son significativas (no se asocian con ARM). La (Tabla 3) nos resume los resultados significativos de las dos tablas previas.

Ahora queremos saber dos cosas: por un lado, ¿cuál de estas 6 variables es la mejor predictora individualmente? y por otro ¿qué combinación de variables me permiten la mejor predicción de ARM?

#### 7. ¿Cuál es mejor? ¿O qué combinación es mejor? (Parte 1)

Para eso utilizaremos ahora los modelos de Regresión Logística múltiples (con varias

variables independientes), pero necesitamos hacer un paréntesis para comentar algo: Abrimos un paréntesis.

La Regresión Logística simple se utiliza comúnmente cuando la variable respuesta es dicotómica (Si/No, o bien 1/0) y además la variable independiente es cuantitativa. Pero es igualmente útil cuando ambas variables son categóricas (y dicotómicas), es decir ambas son Si/No o ambas son 1(uno) o 0(cero).

Recordemos los resultados del punto 5b, donde demostramos que VRS está relacionado significativamente con ARM y esta relación fue probada mediante una prueba de independencia Chi cuadrado. Ahora veremos los resultados de esa asociación, pero puesta a prueba mediante Regresión Logística simple. Si pedimos que en vez de una Prueba Chi cuadrado se realice una Regresión Logística los resultados son equivalentes (Tabla 4).

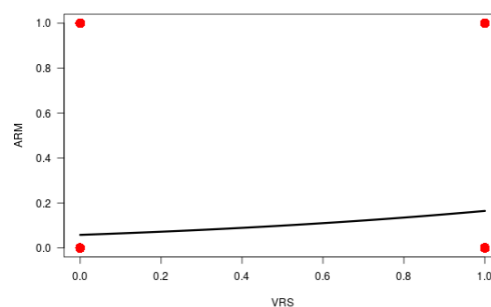
**Tabla 4.** Resultado del modelo de Regresión Logística simple entre VRS (1/0) y ARM (1/0) y su correspondiente valor p.

	Estimador	Error estándar	Valor Wald	Valor p
Ordenada	-2.783	0.4207	-6.6154	<0.001
Pendiente	1.160	0.5068	2.2896	0.022

Y el Odds Ratio estimado es de 3.19, que surge de  $\exp(1.160)$ .

La (Figura 5) representa el modelo ajustado. Como se puede observar, el valor del Odds ratio estimado a partir de la fórmula clásica (de la tabla de doble entrada que usamos en la prueba chi cuadrado) y el estimado en la Regresión Logística es el mismo: 3.19. Asimismo, los valores p obtenidos por ambos métodos poseen sólo una muy pequeña diferencia.

Cerramos el paréntesis



**Figura 5.** Relación VRS (1/0) y ARM (1/0) mediante Regresión Logística. Curva ajustada ( $p=0.022$ )

La pregunta es: ¿cuál de los dos test usamos? La respuesta es: sabiendo que son análisis equivalentes, cuando tenemos sólo una variable independiente usamos la prueba chi cuadrado, ya

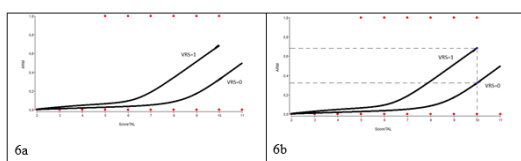
que la tabla (Tabla 1, por ejemplo) es de más fácil interpretación que la (Tabla 4).

Pero para modelos más complejos (que veremos a continuación) sabemos que podemos combinar variables independientes tanto cualitativas como cuantitativas porque la práctica médica podríamos necesitar ambos tipos de variables para predecir la probabilidad de un evento desfavorable.

7 bis. ¿Cuál es mejor? ¿O qué combinación es mejor? (Parte 2)

Aplicamos ahora el Análisis de Regresión Logística múltiple (stepwise), que combina las variables cuantitativas y cualitativas como predictoras, pero además sólo ingresan al modelo, aquellas variables que son significativas y en orden de importancia de las mismas. El resultado de solicitarle al análisis que considere las 6 variables independientes para predecir a ARM se presenta en la Tabla 5. Allí se observa que estadísticamente sólo es necesario predecir el requerimiento de ARM con Score de Tal (numérica) y VRS (Si/No), donde ingresa primero al modelo la variable Score de Tal.

Por su parte, si graficamos la función, ahora nos quedan dos “curvas”, una que corresponde a los pacientes que poseen VRS y otra para los que no lo poseen (Figura 6).



**Figura 6.** Relación entre Score de Tal y VRS como predictores de ARM, mediante Regresión Logística múltiple stepwise. 5a. Curva ajustada. 5b Estimación de valores de ARM en función de dos valores de Score de Tal (10) para pacientes VRS positivos y negativos.

Sólo de la observación gráfica, notamos que la probabilidad más alta de requerir ARM la obtiene un paciente con la combinación de mayor Score de Tal y presencia de VRS (Figura 5).

## Conclusión

El procedimiento sugerido consta de los siguientes pasos ordenados:

1. Búsqueda de las variables planteadas en los objetivos.
2. Armado correcto de la tabla (pacientes en filas y variables en columnas).
3. Clara identificación de la Variable respuesta de aquellas variables independientes.

4. Realización de tablas descriptivas.

5. Realización de pruebas Chi cuadrado de la variable respuesta con las variables independientes cualitativas.

5a. Identificación de las variables significativas de aquellas que no lo son.

5b. Interpretación de los porcentajes condicionados por filas, con la estructura “causa” en filas y “efecto” en columnas.

5c. Interpretación de los Odds ratio.

6. Realización de Regresiones logísticas simples de la variable respuesta con las variables independientes cuantitativas.

6a. Identificación de las variables significativas de aquellas que no lo son.

6b. Interpretación de los Odds ratio

6c. Ante la necesidad de inferir para nuevos casos, calcular la probabilidad esperada mediante la fórmula de la Regresión Logística simple

7. Realización de una Regresión Logística múltiple stepwise de la variable respuesta con todas las variables independientes. Identificación de cuáles son las variables que ingresan al modelo y el orden en que lo realizan.

8. Revisión de todos los resultados anteriores, ya que debe haber una lógica biológica o médica en estos.

Es fundamental que el usuario revise la interpretación de cada uno de los resultados. Los análisis estadísticos son herramientas que ayudan a visualizar mejor los “grises” o las variabilidades presentes de nuestra investigación. Se debe tener presente que los modelos estadísticos y los programas estadísticos que utilizamos para poder realizarlos no tienen la “capacidad” de diferenciar aspectos conceptuales del tema, por lo que es el investigador quien debe estar atento a que los análisis arrojen resultados que inviables biológicamente o que se oponen una manera radical a lo conocido hasta la actualidad. Y si eso ocurre, debemos revisarlo y justificarlo.

Para concluir, los autores esperamos que, mediante este trabajo, junto a la revisión bibliográfica realizada, se logre poner a disposición de los investigadores en salud humana una guía de pasos útiles para analizar las variables cualitativas de sus trabajos.

## Bibliografía

1. Agresti A. Analysis of ordinal categorical data. 2a ed. Hoboken, NJ, Estados Unidos de América: Wiley-Blackwell; 2010.



2. Mangeaud, AP, Bioestadística I, Ferreyra Editor. Córdoba, Argentina. 2014
3. Spiegel MR, Stephens LJ, Estadística. 4° Edición, McGraw-Hill Companies, 2009
4. Siegel, S & N J Castellan. Estadística no paramétrica. 4° Edición. Editorial Trillas;1995
5. Macchi RL. Introducción a la Estadística en Ciencias de la Salud. 3° Edición. Editorial Médica Panamericana. 2020.
6. Henquin,R, Epidemiología y estadística para principiantes. 1° Edición. Editorial Corpus. 2013.
7. Bewick V, Cheek L, Ball J. Statistics review 8: Qualitative data - tests of association. Critical care [Internet]. 2004; 8(1):46–53. Disponible en: <http://dx.doi.org/10.1186/cc2428>
8. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb) [Internet]. 2014;24(1):12–8. Disponible en: <http://dx.doi.org/10.11613/BM.2014.003>
9. Quevedo Ricardi F. The chi-square. Medwave [Internet]. 2011;11(12): e5266–e5266. Disponible en: <http://dx.doi.org/10.5867/medwave.2011.12.5266>
10. Hernández De La Rosa Y, Moreno VJH, Batista Hernández NE. ¿Chi cuadrado o Ji cuadrado? Medicen Electronica. 2017; 21(4). Disponible en: <https://medicentro.sld.cu/index.php/medicentro/article/view/2500/2072>
11. Plackett RL. Karl Pearson and the chi-squared test. Int Stat Rev [Internet]. 1983;51(1):59. Disponible en: <http://dx.doi.org/10.2307/1402731>
12. Christian R, George C. Introducing Monte Carlo methods with R. Springer Science+Business Media; 2010
13. Cerda JC. Rada Odds ratio: aspectos teóricos y prácticos Rev Med Chila. Rev Med Chila 2013; 141: 1329-1335
14. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis [Internet]. 2019;11(Suppl 4): S574–84. Disponible en: <tp://dx.doi.org/10.21037/jtd.2019.01.25>
15. Golan-Tripto I, Goldbart A, Akel K, Dizitzer Y, Novack V, Tal A. Comité de Pediatría Ambulatoria, Colaboradores. Recomendaciones para el manejo de las infecciones respiratorias agudas bajas en menores de 2 años. Arch Argent Pediatr. 2018;119(4):796–801.
16. Golan-Tripto I, Goldbart A, Akel K, Dizitzer Y, Novack V, Tal A. Modified Tal Score: Validated score for prediction of bronchiolitis severity. Pediatr Pulmonol [Internet]. 2018;53(6):796–801. Disponible en: <http://dx.doi.org/10.1002/ppul.24007>
17. Molina P, Bustos S, Valenzuela L, Hidalgo M, Alliu M, Moscoso C, et al. La escala de Tal como test diagnóstico y el diagnóstico clínico como gold standard en el síndrome bronquial obstructivo del lactante Pediatría Atención Primaria. Vol. 37. España; 2008.
18. Giachetto G, Vomero A, Pandolfo S, Notejane M, et al. Validación del score clínico de Tal modificado para la evaluación de severidad en bronquiolitis. Arch Pediatr Urug [Internet]. 2018; Disponible en: <http://dx.doi.org/10.31134/ap.89.s1.5>
19. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform [Internet]. 2018; 116:10–7. Disponible en: <http://dx.doi.org/10.1016/j.ijmedinf.2018.05.006>
20. Heath Gauss DK, Williams DW. Purposeful selection of variables in logistic regression Zoran Bursac. Source Code for Biology and Medicine Source Code for Biology and Medicine [Internet]. 2008;3. Disponible en: <http://dx.doi.org/10.1186/1751-0473-3-17>

21. Mangeaud A, Panigo DE. R-Medic. Un programa de análisis estadísticos sencillo e intuitivo. *Methodo Investig Apl Las Cienc Biol* [Internet]. 2018;3(1):18–22. Disponible en: [http://dx.doi.org/10.22529/me.2018.3\(1\)05](http://dx.doi.org/10.22529/me.2018.3(1)05)
22. IBM Corp. Released. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp. 2011.

