# DOS METODOLOGÍAS ESTADÍSTICAS PARA EL ANÁLISIS DE DATOS CATEGÓRICOS EN REPRODUCCIÓN ANIMAL

**POR:**

**I. I. PATRICIA GUADALUPE ORPINEL UREÑA**

**Tesina presentada como requisito parcial para obtener el grado de**

**Maestría Profesional en Estadística Aplicada**

**Universidad Autónoma de Chihuahua**

**Facultad de Zootecnia y Ecología**

**Secretaría de Investigación y Posgrado**

**Chihuahua, Chih., México** **Diciembre de 2015**

Dos metodologías estadísticas para el análisis de datos categóricos en reproducción animal. Tesina presentada por Patricia Guadalupe Orpinel Ureña como requisito parcial para obtener el grado de Maestría Profesional en Estadística Aplicada, ha sido aprobada y aceptada por:

M. A. Luis Raúl Escárcega Preciado
Director de la Facultad de Zootecnia y Ecología

M. C. Antonio Humberto Chávez Silva
Secretario de Investigación y Posgrado

D. Ph. Pablo Fidel Mancillas Flores
Coordinador Académico

D. Ph. Joel Domínguez Viveros
Presidente

DICIEMBRE 17-2015

Fecha

Comité:

D. Ph. Joel Domínguez Viveros
Dr. Juan Ángel Ortega Gutiérrez
Dr. Nicolás Callejas Juárez



**Patricia Guadalupe Orpinel Ureña**
PERIFÉRICO FRANCISCO R. ALMADA KM. 1, CHIHUAHUA, CHIH., MÉXICO C.P. 31453

## AGRADECIMIENTOS

A Dios por regalarme vida y salud, por permitirme llegar hasta aquí.

A mis padres y hermanos por su apoyo incondicional y por enseñarme el valor de las cosas, por enseñarme a luchar por lo que se quiere, porque todos mis logros, son de ellos también.

A mi esposo por su paciencia y ayuda para que este trabajo se realizara.

A la Universidad Autónoma de Chihuahua que a través de la Facultad de Zootecnia y Ecología me permitió concluir una meta más en mi desarrollo profesional.

Agradezco al D. Ph. Joel Domínguez Viveros por su gran apoyo, su asesoría, sus conocimientos y su paciencia durante el desarrollo de mi trabajo.

Agradezco al Dr. Juan Ángel Ortega Gutiérrez por compartir sus conocimientos a lo largo de mis estudios y por el gran apoyo y enseñanza en la elaboración de este trabajo.

Agradezco al M. P. E. A. Nelson Aguilar Palma por su incondicional disposición para atender mis dudas a lo largo del desarrollo de esta tesina.

A Olga, Lina, Jonathan y Héctor, compañeros de cursos y excelente equipo de trabajo, por brindarme su amistad y apoyo.

## DEDICATORIA

A mis padres:

Alejandrina y Patricio

A mis hermanos:

Erik y Adrián

A mi amor:

Germán

A mi hija:

Valeria, el regalo más hermoso que me ha dado Dios.

## CURRICULUM VITAE

La autora nació el 26 de julio de 1982 en el municipio de Hidalgo del Parral, Chihuahua, México.

| | |
|---|---|
| 2000 – 2005 | Estudios de Licenciatura en el Instituto Tecnológico de Chihuahua en Ingeniera Industrial. |
| 2005 | Residencia Profesional: Labinal Safran Group en la Cd. de Chihuahua, Chih. |
| 2006 – 2008 | Ingeniero de Calidad de Maquinado y Ensamble: AUMA S.A. de C.V. en la Cd. de Chihuahua, Chih. |
| 2008 a la Fecha | Profesor de Asignatura de la Facultad de Ingeniería de la UACH, impartiendo clases a las carreras de Ingeniería Civil, Ingeniería Matemática, Ingeniería Física, Ingeniería en Sistemas, Ingeniería en Geología, Ingeniería en Tecnología de Procesos e Ingeniería en Minas. |

**RESUMEN**

DOS METODOLOGÍAS ESTADÍSTICAS PARA EL ANÁLISIS DE DATOS CATEGÓRICOS EN REPRODUCCIÓN ANIMAL

POR:

I.I. PATRICIA GUADALUPE ORPINEL UREÑA

Maestría Profesional en Estadística Aplicada

Secretaría de Investigación y Posgrado

Facultad de Zootecnia y Ecología

Universidad Autónoma de Chihuahua

Presidente: D. Ph. Joel Domínguez Viveros

Se describen dos diferentes metodologías estadísticas que pueden ser utilizadas para el estudio de variables categóricas (variables de respuesta y variables explicativas). Se analizaron dos conjuntos de datos de reproducción animal, ambos con variable respuesta binaria. En ambos casos el análisis se realizó mediante probabilidades marginales, conjuntas y condicionales; prueba de independencia chi-cuadrada; posteriormente se utilizó regresión logística mediante los procedimientos CATMOD y LOGISTIC de SAS. Se observó que el análisis que genera mayor información para realizar inferencias estadísticas con datos para una variable con respuesta binaria es la regresión logística.

**ABSTRACT**

TWO STATISTICS METHODOLOGIES FOR DATA ANALYSIS CATEGORICAL IN ANIMAL REPRODUCTION

BY:

PATRICIA GUADALUPE ORPINEL UREÑA

Two statistical methodologies that can be used to study categorical variables (response variables and explanatory variables) are described. Two sets of data were analyzed both variable bit answer. In both cases the analysis was done by marginal, joint and conditional probabilities; independence test chi-square was performed; Logistic regression was used subsequently by the SAS LOGISTIC and CATMOD and procedures. It was observed that the greatest statistical analysis gives information response data is binary logistic regression models because with it not only probability and hypothesis testing responses such as if two variables are independent or not obtained.

## CONTENIDO

# LISTA DE CUADROS

# LISTA DE CUADROS DEL APÉNDICE

# INTRODUCCIÓN

Para el análisis estadístico de variables continuas, especialmente las que se ajustan a una distribución normal, existen diversos métodos estadísticos como la regresión lineal, el análisis de varianza y los modelos mixtos, por citar algunos, los cuales se fundamentan en la evaluación de parámetros y estimadores, como la media y la varianza (Infante y Zarate, 2000). Para el análisis de datos categóricos, derivados de una distribución de Bernoulli, binomial o de Poisson, la metodología estadística es diferente, específica y en cierto modo restringida; en el contexto de variables categóricas podemos distinguir tres tipos (Agresti, 2002): 1) binarias, aquellas que sólo pueden tomar dos valores (ejemplo: éxito – fracaso; 0 – 1; Sí – No); 2) ordinales, que pueden tomar múltiples valores, entre los cuales es posible establecer una relación de orden (ejemplo: primero – segundo – tercero; grande – mediano – pequeño); y, 3) nominales, la cuales pueden tomar múltiples valores sin un ordenamiento o jerarquización (ejemplo: azul – rojo – blanco; Chihuahua – Parral – Delicias).

Para el análisis de datos categóricos se han desarrollado algunas pruebas estadísticas, como el caso particular del análisis de independencia a través de chi o ji – cuadrada, o la estimación de frecuencias a través de las tablas de contingencia (Le, 2003); por otro lado, en Software especializados en análisis estadísticos, se han implementado ciertos procedimientos enfocados al análisis de datos categóricos. El programa de análisis estadístico SAS (SAS, por sus iniciales en inglés; SAS, 2001), desarrolló dos procedimientos:

a) El CATMOD, para el análisis con modelos lineales a funciones de respuesta frecuentista.

b) El LOGISTIC, para el análisis de regresión logística mediante máxima verosimilitud.

En los sistemas de producción animal, en el área de reproducción, se generan variables categóricas de tipo binario tales como la tasa de preñez (Cavestany *et al.,* 2001), la dificultad al parto (Cañón, 1986; Silva y Cañón, 2000) y tasa de mortandad (Verde, 2000) que requieren un análisis diferencial por su naturaleza y características.

Con base en lo anterior, el objetivo del presente trabajo fue analizar la información que se generó a partir de dos experimentos en reproducción animal, con base en las pruebas y procedimientos desarrollados para variables binarias.

# REVISIÓN DE LITERATURA

## Desarrollo Histórico del Análisis de Variables Categóricas

El análisis de datos categóricos inició a partir de los trabajos de Birch en 1963, con la demostración en las ecuaciones de verosimilitud para modelos log-lineales, que relacionan las estadísticas mínimas suficientes a sus valores esperados; además de que existe una solución única que satisface al modelo y al enlace de los datos muéstrales (Ato y López, 1996). Posteriormente, Grizzle *et al.* (1969) formularon el enfoque analítico alternativo para el tratamiento de datos categóricos, dentro de la infraestructura del modelo lineal clásico; Goodman (1970) presentó el análisis multivariado de datos cualitativos, a partir de tablas de contingencia; además, determinó la estimación directa de las interacciones entre las variables, y las pruebas indirectas de las hipótesis que conciernen a esas interacciones; Ku *et al.* (1971) presentaron la aplicación de la información mínima discriminante para datos categóricos en tablas de tres y cuatro entradas usando modelos log-lineales; Bishop *et al.* (1975) desarrollaron la estimación de máxima verosimilitud en tablas de contingencia, generalizando la metodología para el caso multidimensional e interpretando los modelos log-lineales en función de la tasa de producto cruzado y el principio de jerarquía; también desplegaron un estadístico de verosimilitud, el cual ayuda a determinar el modelo de mejor ajuste a los datos categóricos. Fienberg y Larntz (1976) presentaron las formas de análisis multivariante para clasificación cruzada de datos categóricos, y las estimaciones de máxima verosimilitud en los modelos log-lineales de datos que provienen de una distribución Poisson o multinominal. McCullagn y Nelder (1983) complementaron el desarrollo del modelo lineal

generalizado; posteriormente, Agresti (2002) recopiló los trabajos realizados en modelos log-lineales utilizando una nueva notación; además, generalizó utilizando las condiciones de colapsibilidad y la tasa de ventaja, mostrando su aplicación a datos nominales y ordinales.

**Distribuciones de Probabilidad para Variables Categóricas**

**Distribución de Bernoulli.** El experimento de Bernoulli, describe el modelo aleatorio más sencillo, con base en las siguientes características:

1) En el experimento sólo se hace un ensayo.
2) En el experimento sólo se admiten dos resultados excluyentes, denominados éxito y fracaso.
3) La probabilidad de éxito es $p(E) = p$
4) La probabilidad de fracaso es $p(F) = 1 – p = q$.
5) X es la variable aleatoria que puede tomar valores de 0 si ocurre fracaso y 1 si ocurre un éxito.

Por consiguiente, $x_i = 1$ si el resultado del *i-ésimo* experimento resulta en éxito y $x_i = 0$ si el resultado del *i-ésimo* experimento resulta en fracaso. Los *n* ensayos de Bernoulli reciben el nombre de proceso de Bernoulli, si los mismos son independientes, con sólo dos resultados posibles, y la probabilidad de éxito permanece constante de ensayo en ensayo (Sahagun, 1994). Los componentes y momentos de la distribución de Bernoulli (Le, 2003; Hines y Montgomery, 2004) son:

$$p(x_1, x_2,....,x_n) = p_1(x_1) = p_2(x_2) =…p_n(x_n);$$

$$p_i(x_i) = p(x_i) = p,\ x_i = 1,\ i = 1,2,...,n;$$

$$p_i(x_i) = p(xi) = (1\text{-}p) = q,\ x_i = 0,\ i = 1,2...,n;$$

$p_i(x_i) = p(x_i) = 0$, en otro caso;

La media: $E(\ _i) = (\quad)\quad (1\ p) = p$

La varianza: $ar(\ _i) = (o^2\ p)\ (i^2\ p)\ - p^2 = p(1-p)$

La función que genera los momentos es:

$M_{xi}(t) = \quad _i(t) = \quad pe^t$

**Distribución binomial.** La distribución binomial se obtiene haciendo *n* pruebas de Bernoulli independientes entre sí, con base en las siguientes características (Sahagun, 1994):

a) *n,* corresponde al número de repeticiones independientes del experimento de Bernoulli.

b) Todas las pruebas deben de tener una probabilidad constante de éxito *p,* y una probabilidad constante de fracaso *q = 1 – p*

c) *x,* es el número de éxitos en las *n* pruebas, consecuentemente *n – x* es el número de fracasos.

La variable *x*, que denota el número de éxitos en *n* ensayos de Bernoulli, tiene una distribución binomial (Le, 2003) dada por *p(x)* = {"*x* éxitos en *n* ensayos"}, donde:

$$p \quad = \ ^n\ p\ ^n \qquad\qquad = \ 1\ 2\ \ldots\ n.$$

La media de la distribución binomial puede determinarse como:

$$E \quad = \ ^n_{=} \ . \frac{n}{n}\ p\ ^n$$

$$= np\ ^n_{=1} \ \frac{(n\ 1)}{(\ \ 1)\ n}\ p\ ^{1\ n}$$

si y = x - 1

$$E(\ ) = np \sum_{y=}^{n\ 1} \frac{(n\ 1)}{y\ \ n\ 1\ y} p^{y}\ ^{n\ 1\ y}$$

$$E(x) = n*p$$

Con base en el planteamiento anterior, la varianza de la distribución binomial puede determinarse como:

$$= \sum_{=}^{n} \frac{^{2}n}{n}\ p\ ^{n}\ \ (np)^{2}$$

$$= n\ n\ 1 \sum_{=}^{n\ 2} p^{2} \frac{n\ 2}{y\ \ n\ 2\ y} p^{y}\ ^{n\ 2\ y}\ \ np\ (np)^{2}$$

De manera que

$$V(x) = n*p*q$$

La función que genera los momentos para la distribución binomial es:

$$M_x(t) = (pe^{t} + q)^{n}$$

Donde e es la base constante exponencial e=2.7183

Un enfoque simple (Hines y Montgomery, 2004) para encontrar la media y la varianza es considerar *x* como la suma de *n* variables aleatorias independientes, cada una con media *p* y varianza *pq*, por lo que

$$x = x_1 + x_2 + \ldots + x_n;$$

Donde,

$$E(x) = p + p + \ldots + p = n*p$$

$$V(x) = p*q + p*q + \ldots + p*q = n*p*q$$

**Distribución geométrica.** El interés en la variable aleatoria *x*, que representa el número de ensayos que tienen que realizarse para que se produzca el primer éxito se le denomina variable aleatoria geométrica. La distribución geométrica parte de una secuencia de ensayos de Bernoulli, con la diferencia de que el número de ensayos no es fijo (Sahagun, 1994; Hines y Montgomery, 2004). El espacio del rango para *x* es Rx = {1,2,3,. . .}, y su distribución está dada por:

$$p \quad = \quad {}^{1}p \qquad\qquad =1\ 2\ \ldots\ n.$$

La media de la distribución geométrica se encuentra a partir de

$$E \quad = \quad _{=1} \ . \ p \ . \quad {}^{1}=p \ . \ \frac{d}{dp} \ _{=1}$$

$$= p \ . \ \frac{d}{dp} \ \ \frac{}{1} \ = \frac{1}{p}$$

La varianza de la distribución geométrica es:

$$= \quad _{=1} \quad {}^{2} . \qquad p \ \ \frac{1}{p} 1^{2} = p \quad _{=1} \quad {}^{2} \quad {}^{1} \quad \frac{1}{p^{2}} = \frac{}{p^{2}}$$

La función que genera los momentos es:

$$M_x(t) = pe^t / (1 - qe^t)$$

**Distribución de Poisson.** La distribución de Poisson se puede desarrollar de dos maneras; el primer desarrollo implica la definición de un *proceso de Poisson;* y el segundo muestra una forma límite de la distribución binomial. Las propiedades del proceso de Poisson son:

a) El número de resultados que ocurren en un intervalo o región específica es independiente del número que ocurre en cualquier otro intervalo o

región del espacio disjunto; de esta forma se dice que el proceso de Poisson no tiene memoria.

b) La probabilidad de que ocurra un solo resultado, durante un intervalo muy corto o en una región pequeña, es proporcional a la longitud del intervalo o al tamaño de la región, y no depende del número de resultados que ocurren fuera de este intervalo o región; y,

c) La probabilidad de que ocurra más de un resultado en tal intervalo corto o que caiga en tal región pequeña es insignificante.

Con una variable aleatoria *x* que sigue una distribución binomial con parámetros *n* y *p* en donde *n* es muy grande y *p* es muy pequeña, la distribución de *x* (cuando *n* tiende a infinito) se aproxima a la distribución llamada Poisson (Sahagun, 1994). La función de probabilidad es:

$$p = \frac{e}{} \qquad = 1\ 2\ \ldots$$

En la distribución binomial, si n es grande y la probabilidad p de ocurrencia de un evento se acerca a cero, de tal manera que q = 1 – p se acerca a 1, el evento se denomina suceso raro o inusual. En la práctica se puede considerar que un evento es raro si el número de ensayos es por lo menos de 50 ($n \geq 50$), mientras que (n*p) < 5. A partir de aquí la distribución binomial se aproxima a la distribución de Poisson con $\lambda$ = np. La distribución de Poisson se representa mediante la siguiente función:

$$p = \frac{e^{np}\, np}{} \qquad = 1\ 2\ \ldots$$

La media y la varianza de la distribución de Poisson es $\lambda$, a partir de los siguientes planteamientos:

$$E = \frac{e^{\lambda}\lambda}{} = \frac{e^{\lambda}\lambda}{1}$$

$$= \lambda e^{\lambda} \; 1\frac{\lambda}{1} \; \frac{\lambda^2}{2} \; \ldots = \lambda e^{\lambda} e^{\lambda} = \lambda$$

De modo similar

$$E^{2} = \frac{^{2} e^{\lambda}\lambda}{} = \lambda^2 \quad \lambda$$

Por lo que

$$V(x) = E(x^2) - [E(x)]^2 = \lambda$$

La función que genera los momentos es:

$$t = e^{c(e^{t} 1)}$$

**Procedimientos de Análisis de Variables Categóricas**

**Prueba chi-cuadrado.** El análisis chi cuadrada ($x^2$) se puede realizar abordando tres criterios (Lipschutz y Lipson, 2001):

1) Para bondad de ajuste o pruebas de homogeneidad; consiste en determinar si los datos de cierta muestra corresponden a cierta distribución poblacional.

2) Para homogeneidad a través de varias muestras cualitativas, que consiste en probar si varias muestras de un carácter cualitativo proceden de la misma población.

3) Para pruebas de independencia, que consiste en comprobar si dos variables cualitativas están relacionadas entre sí (son independientes o no). Por

ejemplo, a un encuestador le interesa saber si el género, los antecedentes étnicos o el rango salarial de una persona son factores relevantes al votar en una elección.

La definición formal de la distribución chi cuadrado es: sean $Z_1$, $Z_2$ … $Z_k$, k distribuciones normales independientes. Entonces $X^2 = {}^2Z_1 + {}^2Z_2 \ldots Z_k{}^2$, se denomina la distribución chi-cuadrado con *k* grados de libertad; el número de grados de libertad (k), puede ser cualquier entero positivo incluyendo 1. Por lo tanto, hay una distribución $X^2$ para cada *k*. La distribución no es simétrica y está sesgada hacia la derecha; sin embargo, para un número grande de *k,* la distribución se acerca a la distribución normal (Lipschutz y Lipson, 2001).

**Tablas de contingencia con prueba de chi-cuadrado ($x^2$).** Las tablas de contingencia se construyen a partir de tablas de doble entrada, donde cada entrada representa un criterio o variable de clasificación; las frecuencias se organizan a través de casillas, que contienen información sobre la relación existente entre ambos criterios. Las tablas de contingencia se utilizan para examinar la relación entre dos variables categóricas, a partir de la prueba de $\boldsymbol{x^2}$. El Cuadro 1 muestra un ejemplo de tabla de contingencia con dos variables: X y Y; donde, X contiene *i* categorías y Y contiene *j* categorías, en total hay ij combinaciones o casillas con información de las frecuencias absolutas. Las distribuciones marginales contienen información acerca de una sola variable e ignoran la asociación que pudiera existir entre las dos variables; en ciertas ocasiones, una de las variables es de respuesta y la otra causal; o en su caso, una variable es aleatoria y la otra fija. En este proceso la distribución conjunta de X e Y no hace sentido; sin embargo, se desea conocer si la distribución de la

Cuadro 1. Tabla de contingencia para el análisis de las frecuencias en cada una de las respuestas posibles

| | | Variable y | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | j |
| Variable x | 1 | $\pi_{11}$ | $\pi_{12}$ | $\pi_{13}$ | $\pi_{14}$ | $\pi_{1j}$ |
| | 2 | $\pi_{21}$ | $\pi_{22}$ | $\pi_{23}$ | $\pi_{24}$ | $\pi_{2j}$ |
| | 3 | $\pi_{31}$ | $\pi_{32}$ | $\pi_{33}$ | $\pi_{34}$ | $\pi_{3j.}$ |
| | i | $\Pi_{i1}$ | $\Pi_{i2}$ | $\Pi_{i3}$ | $\Pi_{i4}$ | $\pi_{ij}$ |

Donde, π representa los conteos en cada casilla; i = número de categorías para la variable x; j = número de categorías para la variable y.

variable de respuesta depende de los niveles de la variable causal. También se pueden obtener de ella probabilidades condicionales, es decir reducir el universo y calcular la posibilidad de que se dé uno de los niveles de una variable dado que ya conocemos el nivel de la otra.

El estadístico $x^2$ propuesto por Pearson, permite probar la hipótesis de independencia a través de los dos criterios de clasificación utilizados (dos variables categóricas); compara las frecuencias observadas u obtenidas con las frecuencias esperadas, que corresponde a las frecuencias que teóricamente debería haber encontrado en cada casilla, si los dos criterios de clasificación fueran independientes. Las frecuencias esperadas se estiman de la siguiente manera:

$$(\text{frecuencia esperada})_i = \frac{\text{total de fila i (total de columna )}}{\text{nº total de casos}}$$

Bajo la condición de independencia, la frecuencia esperada de una casilla se obtiene dividiendo el producto de las frecuencias marginales correspondientes a esa casilla (su total de fila y su total de columna) por el número total de casos. Obtenidas las frecuencias esperadas para cada casilla, el estadístico $x^2$ se obtiene como la relación entre la diferencia cuadrada de la sumatoria de las frecuencias observadas ($n_{ij}$) y las frecuencias esperadas ($m_{ij}$) entre las frecuencias esperadas a través de todas las celdas

$$^2 = \frac{(n_i \ m_i)^2}{m_i}$$

$_i$

El estadístico $x^2$ tiende a cero cuando las variables sean completamente independientes; por consiguiente, $x^2$ se incrementa cuanto mayor sea la discrepancia entre las frecuencias observadas y las esperadas. El estadístico $x^2$ sigue el modelo de distribución de probabilidad $x^2$, con (i-1)*(j-1) grados de libertad.

Para que las probabilidades de la distribución $x^2$ constituyan una buena aproximación a la distribución del estadístico $x^2$, conviene que se cumplan algunas condiciones; por ejemplo, que las frecuencias observadas no sean demasiado pequeñas; si existen frecuencias esperadas pequeñas o cercanas a cero, estas no deben superar el 20 % del total de frecuencias esperadas; en el caso de que sea mayor el estadístico de Pearson debe ser interpretado con cautela.

**Modelos para Variables Categóricas**

Los modelos de respuesta discreta son la herramienta estadística apropiada para modelar el comportamiento de una variable dependiente de naturaleza discreta a partir de un conjunto de variables independientes que pueden ser tanto discretas como continuas; estos modelos son un caso particular de los ‘‘ odelos Lineales Generalizados” introducido por Nelder y Wedderburn (1972). Los modelos de respuesta discreta pueden clasificarse en modelos de respuesta binaria (dos categorías) y modelos de respuesta múltiple (más de dos categorías). Si existe un orden natural en las categorías, entonces es un modelo de respuesta ordenada.

Los modelos de respuesta binaria tienen una variable dependiente (***Y***) dicotómica que puede tomar valores: 0 y 1. Generalmente se asocia el valor de

a la “ausencia” al “fracaso” o a una respuesta negativa, y 1 a la “presencia” el “é ito” o una respuesta afirmativa. La variable ***Y*** sigue una distribución de Bernoulli de parámetro p(0<p<1); puesto que existe una serie de valores independientes “X” (predictoras) del comportamiento de Y, lo propio es considerar la distribución de Y en cada valor observado de X “Y( )”. Se tiene que es también una Bernoulli de esperanza p(x) y varianza p(x)[1-p(x)]. El objetivo será la construcción de un modelo para Y(x).

**Modelo lineal.** Suponiendo R variables independientes, N observaciones en cada una y el vector que contiene las observaciones de cada variable para el i-ésimo individuo. El modelo de probabilidad lineal, que origina el modelo de regresión lineal

$$Y_{\ i} = a \quad b_{\ i} \quad e_{\ i} \text{ para } i = 1 \ldots N;\ j = 1 \ldots R$$

Es de la forma:

$$E\ Y\ X_i = {}_{\ i} = p_{\ i} = a \qquad b_{\ i} \text{ para } i = 1 \ldots N;\ j = 1 \ldots R$$

Este modelo presenta problemas de normalidad, homocedasticidad (varianza de la respuesta no constante sobre los valores de x), la posibilidad de obtener valores de la probabilidad por debajo de cero y por encima de uno, la subestimación del parámetro $R^2$ y, sobre todo, el hecho de que aumentos iguales en las variables explicativas originen aumentos iguales en la probabilidad de respuesta. Esta última situación no es en absoluto realista, ya que en general esta dependencia no será lineal.

Todos los problemas presentados hacen que estos modelos no sean tan utilizados y, en su lugar, se prefieren los modelos no lineales, que vienen a corregir dichos problemas. Los modelos no lineales buscan que:

$$p_{\ i} = \quad a \qquad b_{\ i} \qquad i=1 \ldots N \ (=1 \ldots R)$$

Es decir:

$$^{1}\ p_{\ i} = a \qquad b_{\ i} \quad i=1 \ldots N \ (=1 \ldots R)$$

La elección de esa función F determina el modelo considerado. Uno de los modelos más destacados es el logit.

**Modelo logit.** Siguiendo con la notación anterior, el modelo de regresión logística es de la forma:

$$p_{\ i} = \frac{e^{(a \quad b_{\ i})}}{1 \ e^{(a \quad b_{\ i})}} \quad i=1 \ldots N \ (=1 \ldots R)$$

O su equivalente:

$$\ln \frac{p(\ _i)}{1\ p(\ _i)} = a \qquad b_{\ i} \quad i=1 \ldots N \ (=1 \ldots R)$$

Las estimas de este modelo no se salen del rango [0,1] como ocurría en el lineal. Además, las rectas Y=0 e Y=1 son asíntotas horizontales y la tasa de cambio en $p(x_i)$ es variable.

**Modelo de regresión logística binaria.** El modelo de regresión logística binaria es de interés desde el punto de vista de la reproducción animal. La formulación del modelo logit viene expresado a partir de:

$$\ln \frac{p(\ _i)}{1\ p(\ _i)} = a \qquad b_{\ i} \quad i=1 \ldots N \ (=1 \ldots R)$$

donde k es el número de variables independientes, n el número de observaciones en cada variable, $_i=(_{i1} \ldots \quad _{iR})$ el vector de observaciones de cada variable para el i-ésimo individuo y el cociente $\frac{p(\ )}{1-p(\ )}$ representa la ventaja de respuesta Y=1 para los valores observados de las variables independientes. El caso más sencillo de modelo logístico es aquél en el que se tiene una única variable independiente continua, esto es:

$$\ln\frac{p(_i)}{1\ p(_i)} = a\ b$$

Las principales características de la curva de respuesta en el caso de una variable con un sólo parámetro (b) son que la curva tiene forma de S y está acotada dentro del intervalo de valores [0,1], donde las rectas Y=0 e Y=1 son asíntotas horizontales. Su crecimiento es monótono, pudiendo ser creciente si (b>0) o decreciente (si b<0). Por tanto, con b>0 la probabilidad de respuesta tenderá a uno cuando y cero cuando . La situación se invierte si b<0. Si b=0 la curva es en realidad una recta e Y es independiente de x. La tasa de cambio en p(x) por cada unidad de cambio en x es variable, ya que viene dada por la pendiente de la recta tangente a la curva.

Si se tiene alguna variable independiente categórica, es necesario definir una serie de variables nuevas, artificiales, que servirán para poder pasar de una variable categórica con k categorías a k-1 variables indicadoras de la presencia de cada categoría, por separado. Dichas variables de diseño, conocidas como variables “dummy” son introducidas en el modelo como variables continuas, tal como se explica a continuación:

Para crear k-1 variables de diseño, asociadas a una variable con k categorías, se tienen dos métodos posibles: el método parcial y el método marginal:

a) El método parcial consiste en elegir una categoría de referencia dentro de las k posibles, construir para cada una de las restantes una variable que valga 1 en la categoría considerada, y 0 en el resto. Por ejemplo, si tenemos una variable con las categorías “ba o” “medio” y “alto” se puede elegir “ba o” como categoría de referencia y crear dos variables de diseño: una que valga uno con presencia de la categoría “medio”, y 0 en los otros dos casos; una segunda variable que valga 1 cuando se presente la categoría “alto”, y 0 en los otros dos casos; estas dos variables son las que se modelan.

b) El método marginal es similar al parcial, salvo que todas las variables toman el valor -1 cuando se da la categoría de referencia, en lugar de 1. Lo común es utilizar el método parcial que permite interpretar los parámetros en términos de cocientes de ventajas de forma sencilla.

**Modelos lineales a través de CATMOD del SAS.** Dentro del paquete de análisis estadístico SAS (2001) existen varios procedimientos para el análisis de datos categóricos mediante procedimientos logísticos. El CATMOD es un procedimiento que se ajusta a funciones de datos categóricos, facilitando su análisis con regresión, análisis de varianza, modelos lineales, modelos log lineales, regresión logística y análisis de medidas repetidas. La estimación de máxima verosimilitud es utilizada para el análisis de logística y logística generalizada; el análisis de mínimos cuadrados es usado para ajustar modelos a funciones con otras respuestas.

El procedimiento CATMOD arroja la estimación de máxima verosimilitud para la regresión logística, incluyendo el análisis logístico para respuestas dicotómicas y análisis generalizado logístico para respuestas policotómicas. Proporciona la estimación de mínimos cuadrados de otras funciones de respuesta, media, además calcula y analiza las funciones de otras respuestas que se pueden generar a partir de proporciones correspondientes de una tabla de contingencia. Para este procedimiento todas las variables explicatorias deben ser clasificatorias (SAS, 2001).

Para utilizar este procedimiento, las variables respuesta deben ser dicotómicas. Sea $x_1$, $x_2$… $_v$ el conjunto de variables explicativas, por simplicidad suponemos que Y toma valores 0 y 1 con y por tanto π (Stokes e*t al.*, 2000). El CATMOD analiza datos que pueden ser representados por una tabla de contingencia; se asume que las frecuencias (ij) en la tabla siguen una distribución multinomial donde la muestra se obtiene al azar de una población (Le, 2003). Para cada muestra i, la probabilidad de que la j-ésima respuesta $\pi_{ij}$ es estimada por la proporción $p_{ij} = n_{ij}|n_i$. El vector (p) de todas las proporciones es transformado en un vector de la función = (p). Si π indica ue el vector de la probabilidad verdadera para la tabla entera, entonces la función de la probabilidad verdadera se representa, por (π). El modelo indica que la excepción asintótica ( ) es igual a la función de probabilidad verdadera, es decir, igual a la matriz de constantes fijas (X) por el vector de parámetros estimados.

$$E_A = \pi = X$$

El CATMOD utiliza los métodos de estimación de máxima verosimilitud y mínimos cuadrados ponderados. El método de máxima verosimilitud que estima los parámetros del modelo lineal así como el máximo de los valores de la función de verosimilitud multinomial conjunta de la respuesta. El método de mínimos cuadrados ponderados, estima la suma de los residuales. La sintaxis general del procedimiento CATMOD se muestra en el Cuadro 2.

**Regresión logística mediante máxima verosimilitud y procedimiento LOGIT de SAS.** La regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica, epidemiológica y genética, de ahí su amplia utilización. El objetivo que resuelve esta técnica es modelar como influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos; también se utiliza para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso, con más de dos categorías (politómico).

Para los modelos de respuesta binaria, la respuesta (y) de un individuo o unidad experimental puede tomar uno, de dos valores posibles, los cuales se pueden expresar con y = 1 si una enfermedad está presente, ó y = 0 si no lo está. Con la suposición ue X es un vector de variables e plicativas y $\pi = p(y=1|x)$ es la probabilidad de respuesta a modelar (SAS,2001). El modelo logístico lineal tiene la forma:

$$\log(\pi) = (\pi/(1-\pi)) = \alpha \quad ´x$$

Cuadro 2. Sintaxis del procedimiento CATMOD en el programa de análisis estadístico SAS

| Instrucción | Opciones o complementos |
|---|---|
| PROC CATMOD | <options>; |
| DIRECT | <variables>; |
| MODEL | response-effect = design-effects ; |
| CONTRAST | 'label´ row-description </option>; |
| BY | variables; |
| FACTORS | factor-description; |
| LOGLIN | effects; |
| POPULATION | variables; |
| REPEATED | factor-description; |
| RESPONSE | function; |
| RESTRICT | parameter = value; |
| WEIGHT | variable; |

*Adaptado del manual de procedimientos del programa para análisis estadísticos SAS (2001)

Donde α es el parámetro de intersección y = ( _1… _s )´ es el vector de parámetros dependiente. El procedimiento LOGISTIC, modela la probabilidad más baja de los niveles de respuesta, a partir de técnicas de regresión; sin embargo, la metodología de la regresión lineal no es aplicable, dado la naturaleza de la variable respuesta. La regresión logística tiene relación con el parámetro de cuantificación de riesgo, conocido como “odds ratio”; el odds asociado a un suceso, es el cociente entre la probabilidad de que ocurra (p) un evento, frente a la probabilidad de que no ocurra (1-p).

Odds ratio = p / (1-p)

La noción que se está midiendo es parecida al denominado riesgo relativo, el cual corresponde al cociente de la probabilidad de que aparezca un suceso cuando está presente el factor, respecto a cuándo no lo está.

En la ecuación de regresión hay un factor dicotómico (tipo uno *vs* tipo dos), el coeficiente b de la ecuación para ese factor está directamente relacionado con el odds ratio (OR) de usar tipo uno o tipo dos.

OR = exp(b).

El exp(b) es una medida del riesgo que representa poseer el factor correspondiente, con respecto a no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes.

Cuando la variable es numérica, es una medida que cuantifica el cambio en el riesgo cuando se pasa de un valor del factor a otro, permaneciendo constantes el resto de las variables. Así el odds ratio, que supone pasar de $X_1$ a $X_2$, siendo b el coeficiente correspondiente en el modelo logístico: OR = exp [b(X2-X1)]. Se trata de un modelo en el que el aumento o disminución del

riesgo, al pasar de un valor a otro del factor, es proporcional al cambio, es decir a la diferencia entre los dos valores, pero no al punto de partida. Por ejemplo, con el modelo logístico, el cambio en el riesgo de muerte a través de la edad del individuo, es el mismo cuando pasamos de 40 a 50 años que cuando pasamos de 80 a 90. Cuando el coeficiente b de la variable es positivo, existe un odds ratio mayor que 1, y corresponde por tanto a un factor de riesgo. Por el contrario, si b es negativo el odds ratio será menor que uno, y se trata de un factor de protección. En la mayoría de los estudios se tienen varias variables, y por medio de modelación se puede hacer un análisis más eficiente, dado que generalmente se quiere describir los efectos de varias variables explicativas en una o más variables de respuesta. Para esto existen los modelos lineales generalizados, los cuales tienen tres componentes:

a) Un componente aleatorio, que corresponde a la distribución de probabilidad de la variable de respuesta.

b) Un componente sistemático, el cual especifica una función lineal de las variables explicativas que se usa como predictor.

c) Un enlace, el cual describe la relación funcional entre el componente sistemático, y el valor esperado del componente aleatorio (link function, por su origen del inglés).

Para explicar el comportamiento de una variable dependiente binaria se puede usar un modelo logit de la forma:

$$Y = f(\beta_1 \quad \beta_2 x_2 \quad \ldots \quad \beta_k x_k) + u;$$

donde, f es la función logística f(z) = exp〖(z)〗 / (1+ exp(z)). Por lo tanto, E[Y] = P(Y=1) = exp〖( 1 2 x2 … k xk)〗 / (1+ exp( 1 2 x2 … k xk)). La estimación de modelos logit se realiza mediante el método de máxima verosimilitud; en estos modelos no resulta posible interpretar directamente las estimaciones de los parámetros ya ue son modelos no lineales. Si el estimador es positivo, significa que los incrementos en la variable asociada causan incrementos en P(Y=1), aunque se desconoce la magnitud de los mismos. Por el contrario, si el estimador muestra un signo negativo, ello supondrá que incrementos en la variable asociada causarán disminuciones en P(Y=1). En el modelo Logit se usan otros dos conceptos para profundizar más en la interpretación de los estimadores:

1) Se llama “odds” al cociente de probabilidades: (P(Y=1)) / (1-P(Y=1)) = exp( 1 2 x2 … k xk).

   A partir de los logaritmos neperianos, se obtiene una expresión lineal para el modelo:

   Logit [P(Y=1) ] = ln((P(Y=1)) / (1-P(Y=1))) = 1 2 x2 … k xk

El estimador del parámetro 2 se podrá interpretar como la variación en el término logit (logaritmo neperiano del cociente de probabilidades) causada por una variación unitaria en la variable $X_2$ (suponiendo constantes el resto de las variables explicativas).

2) Cuando se hace referencia al incremento unitario en una de las variables explicativas del modelo, aparece el concepto de odds-ratio como el cociente entre los dos odds asociados (el obtenido tras realizar el incremento y

el anterior al mismo). Con el supuesto de un incremento unitario en la variable $X_i$, se tiene: Odds-ratio = (Odds 2)/(Odds 1) = exp〖( $_i$)〗. Un odds-ratio cercano a uno, señala que cambios en la variable independiente asociada, no tendrán efecto alguno sobre la variable dependiente Y. La sintaxis general del procedimiento logistic se describe en el Cuadro 3.

Cuadro 3. Sintaxis del procedimiento LOGISTIC en el programa de análisis estadístico SAS

| **Instrucción** | **Opciones o complemento** |
|---|---|
| PROC LOGISTIC | <options>; |
| BY | variables; |
| CLASS | variable<(v-options)> <variables)> |
| CONTRAST | <effect values></options>; |
| EXACT | <'label' - Intercept><effects></options>; |
| FREQ | variable; |
| MODEL | events/trials=<effects></options>; |
| MODEL | <(variable_options)> ; |
| OUTPUT | <OUT=SAS-data-set>; |
| SCORE | <options>; |
| STRATA | effects</options>; |
| UNITS | <…independent=listk></option>; |
| WEIGHT | variable</option>; |

*Adaptado del manual de procedimientos del programa para análisis estadísticos SAS (2001)

# MATERIALES Y MÉTODOS

## Estudio de Caso Uno

Se analizó la información publicada por Jaramillo (2003), la cual correspondió a un experimento de inseminación artificial en ovejas, con el objetivo de validar el posible efecto del dilutor de semen, en la tasa de gestación. Se inseminaron 80, analizando la variable respuesta de gestación (si - no), de carácter categórico y binaria, a partir de las variables independientes de tipo de dilutor de semen con dos niveles (DIL; DIL1 y DIL2), y el semental de origen del semen, también con dos niveles (SEM; SEM1 y SEM2). Primero, se estimaron las posibles diferencias en probabilidades (marginal, condicional y conjunta) de gestación entre SEM y DIL, con tablas de doble entrada, como la ilustrada en el Cuadro 4; en conjunto, dentro de SEM (fijando SEM1 y SEM2), y de DIL (fijando DIL1 y DIL2), se analizó las probabilidades de gestación. Posteriormente, se analizó la posible asociación de las variables independientes con la variable respuesta, con base en las pruebas: Chi - cuadrada y exacta de Fisher; a partir de la hipótesis nula de independencia, los valores teóricos se obtuvieron a partir de las frecuencias marginales (Ato y López, 1996). Los análisis se realizaron con el procedimiento FREQ del programa para análisis estadístico SAS (SAS, 2001).

Con el procedimiento CATMOD (SAS, 2001), utilizando la metodología de máxima verosimilitud y de mínimos cuadrados, se estimaron las probabilidades de la j - esima respuesta con base en el modelo lineal: gestación = SE DIL; donde la probabilidad de é ito (1; πi ) estuvo definida como: π

Cuadro 4. Tabla de doble entrada para analizar la relación entre la variable independiente de dilutor, con la variable respuesta de gestación

| Dilutor | Gestación | | Total |
|---|---|---|---|
| | No | Si | |
| 1 | 7 | 33 | 40 |
| 2 | 9 | 31 | 40 |
| Total | 16 | 64 | 80 |

[1]Todos los conteos por celda cumplen con ser mayores de 5, lo cual hace válida la prueba de Chi cuadrada.

=P(Y=1|$X_1$,$X_2$); en contra parte, la probabilidad de fracaso (0) tuvo el planteamiento: 1 – π =P(Y= | $X_1$,$X_2$).

Los estadísticos derivados del procedimiento fueron: $z=b_j/s_j$, donde $s_j$ es el error estándar (la raíz cuadrada de la cuasivarianza muestral) de $b_j$. El software SAS utiliza $z^2$, el estadístico de Wald, que sigue una distribución normal y el cual contrasta la hipótesis de que un coeficiente aislado es diferente de cero; su valor para un coeficiente oncreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar, la obtención de significación indica que dicho coeficiente es diferente de cero y vale la pena su conservación en el modelo, los odds ratio para dar respuesta a preguntas de interés como, por ejemplo, ¿cuánto más probable es tener gestación si se utiliza el dilutor 2 en relación a utilizar el dilutor 1? y chi-cuadrado de cociente de verosimilitudes para comprobar la bondad de ajuste del modelo (la cercanía de los valores predichos por el modelo a los observados); con los cuales se evaluaron las hipótesis nula de que todos los coeficientes de regresión logística son : Ho ( $_j$=0) versus de que al menos uno es diferente de cero Ha ( $_j\neq$ ). Para la transformación lineal se utilizó la función *logit* para extender el modelo de regresión lineal

$$Y= \quad {}^{v}_{=1} \quad X \quad a \ln \frac{\pi}{1-\pi} = \alpha \quad {}^{v}_{=1} \quad X \quad \pi= \frac{e^{\alpha \ {}^{v}_{=1} X}}{1 \ e^{\alpha \ {}^{v}_{=1} X}}$$

y así

$$1-\pi= \frac{1}{1 \ e^{\alpha \ {}^{v}_{=1} X}} .$$

Los parámetros a estimar son α y los coeficientes de regresión logística ( $_j$), para ello se consideró la función de verosimilitud:

$$L=\prod_{i=1}^{n} P\left(\frac{Y_i}{X_{i1}\ldots X_{iv}}\right)=\prod_{i=1}^{n}\left(\frac{e^{\sum_{=1} X}}{1\ e^{\sum_{=1} X}}\right)^{Y_i}\left(\frac{1}{1\ e^{\sum_{=1} X}}\right)^{1\ Y_i}$$

Que proporcionaron las estimaciones máximo verosímiles mediante un proceso iterativo.

Por otro lado, se planteó un modelo de regresión logística $p=\frac{1}{1\ e^{-(\alpha\ \ \alpha_1\ 1\ \alpha_2\ 2)}}$

Los análisis se realizaron con el componente para regresión logística del SAS (PROC LOGIST; SAS, 2001); este componente utiliza la interpretación de los parámetros con base en su signo, si el estimador es positivo significa que incrementos en la variable asociada causan incrementos en P(Y=1) aunque se desconoce la magnitud de los mismos, por el contrario si el estimador muestra un signo negativo, ello supondrá que incrementos en la variable asociada causarán disminuciones en P(Y=1); así como las funciones de enlace CLOGLOG|LOGIT|PROBIT, las cuales relacionan el componente aleatorio con la parte sistemática.

En la fórmula del modelo se tiene una serie de coeficientes que son los parámetros del mismo. Además a partir de ellos se calcularon los “cocientes de venta as” para el caso de la respuesta Y=1 dados dos valores distintos $_1$ y $x_2$ del predictor, se calcularon con:

$$_{12}=\frac{\frac{p(\ _1)}{1\ p(\ _1)}}{\frac{p(\ _2)}{1\ p(\ _2)}} \qquad _{21}=\frac{\frac{p(\ _2)}{1\ p(\ _2)}}{\frac{p(\ _1)}{1\ p(\ _1)}}$$

De esta manera se interpretaron los parámetros del modelo en términos de cocientes de ventajas.

**Estudio de Caso Dos**

En este apartado se utilizó la información generada por Contreras (2014), a partir de un proyecto de reproducción en bovinos. Se utilizaron datos de lavados de tres razas utilizadas como donadoras sometidas a protocolos de superovulación para la colecta de embriones de diferentes ranchos del estado de Chihuahua, México. El estudio tuvo una duración de tres años, del 2011 hasta el 2013, periodo durante el cual las hembras donadoras fueron sometidas a los tratamientos. Las características de las donadoras fueron vacas adultas de las razas Angus, Charolais y Hereford donde se evaluó los posibles efectos de la variable RAZA y la variable PROTOCOLO en la variable respuesta EMBRIONES. Las vacas del protocolo 1 se sometieron a un protocolo basado en la Hormona Foliculoestimulante (FSH porcina-Folltropin-V) y las vacas del protocolo 2 al protocolo completo más Hormona Liberadora de Gonadotropinas-Fertagyl. La base de datos estuvo conformada por 892 observaciones de presencia o no de embriones distribuidas a través de las variables independientes: 1) raza, con tres niveles (RAZA; R1, R2 y R3); y, 2) protocolo, con dos niveles (PROTOCOLO; P1 y P2). Al igual que en el caso uno se comenzó estimando las posibles diferencias en probabilidades (marginal, condicional y conjunta) de presencia de embrión entre RAZA y PROTOCOLO, a partir del planteamiento de tablas de doble entrada; conjuntamente, dentro de RAZA fijando P1 y P2, y dentro de PROTOCOLO fijando R1, R2 y R3, también se analizó las probabilidades de presencia de embrión. Posteriormente, se analizó la posible asociación de las variables independientes con la variable respuesta, con base en las pruebas: Ji - cuadrada y exacta de Fisher; a partir

de la hipótesis nula, los valores teóricos se obtuvieron a partir de las frecuencias marginales (Ato y López, 1996). Los análisis se realizaron con el procedimiento FREQ del programa para análisis estadístico SAS (SAS, 2001).

Con el procedimiento CATMOD (SAS, 2001), utilizando la metodología de máxima verosimilitud y de mínimos cuadrados, se estimaron las probabilidades de la j - ésima respuesta con base en el modelo lineal: EMBRIÓN = RAZA + PROTOCOLO; donde la probabilidad de é ito (1; πi ) estuvo definida como: $\pi = P(Y=1|X_1,X_2)$; en contra parte, la probabilidad de fracaso (0) tuvo el planteamiento: $1 - \pi = P(Y= |X_1,X_2)$.

Los estadísticos derivados del procedimiento fueron similares a los ya descritos en el caso uno, puesto que se proponen las mismas metodologías en ambos casos.

# RESULTADOS Y DISCUSIÓN

## Estudio de Caso Uno

Primero se obtuvieron las probabilidades marginales: probabilidad de gestación en macho uno y dos igual a 0.8 respectivamente; probabilidad de gestación del dilutor uno 0.825 y dilutor dos 0.775; estos resultados dan una idea inicial de que no hay diferencia alguna en macho para la probabilidad de gestación, y la diferencia en dilutor es mínima ($p>0.05$). Posteriormente, se obtuvieron las probabilidades condicionales de gestación para cada dilutor fijando el nivel uno de la variable macho, obteniendo 0.85 para el dilutor uno y 0.75 para el dilutor dos, con ellas se pueden calcular de forma empírica los odds ratio; de igual forma, las probabilidades resultantes de gestación para el nivel dos de la variable macho fue igual a 0.8 para ambos dilutores respectivamente. Las probabilidades de gestación a través de macho, fijando los niveles de la variable dilutor, fueron 0.85 para el macho uno y 0.80 para el dos; para el dilutor dos, se obtuvieron probabilidades de 0.75 en el macho uno, y 0.8 en el dos.

Los resultados mostraron que las variables macho y dilutor no tienen relación con la variable respuesta gestación, dado la prueba de chi cuadrado en cada variable con la variable respuesta, gestación. Los valores de ***p*** para las pruebas de la variable independiente dilutor contra la variable respuesta gestación, examinado macho uno, fueron 0.4292 y 0.999, respectivamente; en ambos casos no se rechaza la hipótesis nula ($p > 0.05$) de independencia; de igual forma, los valor de ***p*** para las pruebas dilutor - gestación, a través de macho dos fueron 0.99 y 0.67, con la decisión de no rechazar la hipótesis nula

(p > 0.05) de independencia. El valor de ***p*** para la prueba macho - gestación, en dilutor uno fue 0.6773 y en dilutor dos fue de 0.70, en ambos casos se acepta la hipótesis nula de independencia. En algunas celdas se tiene un reducido número de observaciones, lo cual implica que las aproximaciones asintóticas tipo Chi-cuadrado deberían interpretarse con reservas. En el Cuadro 5 se muestra la prueba chí cuadrada para las variables macho y dilutor.

Con el procedimiento CATMOD se obtuvieron las estimaciones máximo verosímiles mediante un proceso iterativo. Los parámetros y estimaciones fueron:

Parámetro uno = -1.3937

Estimación uno = $-3.99E^{-7}$

Estimación 2=-0.1569

Los valores de las estimaciones representan el cambio diferencial para la variable independiente. El valor $-3.99E^{-7}$ es el cambio diferencial para el macho 1 (para el macho 2 –($-3.99E^{-7}$)= $3.99E^{-7}$; el valor estimado -0.1569, representa el cambio diferencial para el dilutor uno, mientras que para dilutor dos 0.1569 (–(-0.1569)). En el Cuadro 6 se muestran éstos resultados. El ajuste del modelo completo se comprobó mediante el contraste de hipótesis de que todos los coeficientes de regresión logística son cero, mediante un estadístico chi-cuadrado. El estadístico de chi-cuadrado, derivado del cociente de verosimilitudes y de utilidad para comprobar la bondad de ajuste del modelo (la cercanía de los valores predichos por el modelo a los observados) fue de 0.32, con un valor de p de 0.57; por consiguiente no se rechaza la hipótesis nula, de que el modelo es acertado. Se calcularon los valores $_{ij}$ (estimación de la proba-

Cuadro 5. Pruebas chi-cuadrado para cada variable y para el modelo

| Fuente de variación | Grados de libertad | Chi cuadrado | Pr > ChiSp |
|---|---|---|---|
| Término independiente | 1 | 24.55 | <.0001 |
| Macho | 1 | 0 | 1 |
| Dilutor | 1 | 0.31 | 0.5770 |
| Ratio de probabilidad | 1 | 0.32 | 0.5733 |

[1]La probabilidad del término independiente es menor a 0.0001 lo cual indica significancia (P<0.05).
[2]Las variables macho y dilutor no son significantes (P>0.05).

Cuadro 6. Análisis de estimadores de máxima verosimilitud en el estudio de caso uno

| Parámetro | Estimador | Error Estándar | Chi-cuadrado | Pr>chiSq |
|---|---|---|---|---|
| Término independiente | -1.3937 | 0.2813 | 24.55 | <0.0001 |
| Macho 1 | -399E-19 | 0.2801 | 0.00 | 1.0000 |
| Dilutor 1 | -0.1569 | 0.2813 | 0.31 | 0.5770 |

[1]El nivel de la variable Macho 1 es no significante (P>0.05).
[2]El nivel de la variable Dilutor 1 es no significativo (P>0.05).

bilidad de que se produzca la primera respuesta, que no se dé la gestación) en los niveles i y j de las correspondientes variables e plicativas: $\Theta_{11}$= 0.3245. También se contrastó la hipótesis de si los distintos coeficientes de regresión logística son significativos o no ($H_0$: $_j$=0); SAS utiliza $z^2$, el estadístico de Wald, que sigue una distribución Chi-cuadrado con un grado de libertad.

**Estudio de Caso Dos**

Primero se obtuvieron las probabilidades marginales: probabilidad de la existencia de embriones en el protocolo uno fue 0.84; probabilidad de la existencia de embriones en protocolo dos 0.53; éstos resultados dan una idea de que existe diferencia entre utilizar el protocolo uno y dos para la existencia de embriones. Las probabilidades marginales de la existencia de embriones debido a la raza fueron 0.53 para la raza uno, 0.47 para la dos y 0.65 para la tres; éstos resultados dan una idea del parecido que hay en la existencia de embriones debido a la raza aunque deben de interpretarse con cuidado porque cada una de ellas fue calculada sin tomar en cuenta el otro factor. Posteriormente, se obtuvieron las probabilidades de gestación para cada protocolo fijando el nivel uno de la variable raza, obteniendo 0.48 para el protocolo uno y 0.55 para el dos y aquí se observó que dichas probabilidades son muy parecidas; de igual forma, las probabilidades resultantes de existencia de embriones para el nivel dos de la variable raza fueron 0.5 para el protocolo uno y 0.47 para el dos, las probabilidades de existencia de embriones para el nivel de raza tres fueron 0.67 para el protocolo uno y 0.63 para el protocolo 2. Las probabilidades de existencia a través de raza, fijando el nivel uno de la

variable protocolo, fueron de 0.48 para la raza uno, de 0.5 para la dos y 0.67 para la tres; fijando el protocolo dos, se obtuvieron probabilidades de 0.55 en la raza uno, 0.47 en la raza dos y 0.63 en la tres. Estos resultados muestran que existen probabilidades puntualmente parecidas para los diferentes niveles de las variables, lo cual sugiere profundizar en el análisis estadístico para determinar si los diferentes niveles son estadísticamente iguales.

Los resultados mostraron que las variables protocolo y dilutor no tienen relación con la variable respuesta gestación, dado la prueba de chi-cuadrado en cada variable con la variable respuesta, existencia de embriones. Los valores de ***p*** para las pruebas de la variable independiente protocolo contra la variable respuesta existencia de embriones, en la raza uno, fue de 0.759, 0.8870 y 0.7197 para razas dos y tres, respectivamente; en los tres casos no se rechaza la hipótesis nula ($p > 0.05$) de independencia; de igual forma, los valor de ***p*** para las pruebas protocolo – existencia de embriones, fueron 0.1075 para el protocolo uno y 0.1024 para el dos, con la decisión de no rechazar la hipótesis nula ($p > 0.05$) de independencia.

Con el procedimiento CATMOD se obtuvieron las estimaciones máximo verosímiles mediante un proceso iterativo. Los parámetros y estimaciones fueron:

Parámetro uno = -0.1513

Estimación uno = 0.0864

Estimación 2=0.3776

Estimación tres= 0.1195

Los valores de las estimaciones representan el cambio diferencial para la variable independiente. El valor 0.0864 es el cambio diferencial para el protocolo 1 (para el protocolo 2 –(0.0864)= -0.0864; el valor estimado 0.3766, representa el cambio diferencial para la raza uno, mientras que para raza dos es 0.3776. El ajuste del modelo completo se comprobó mediante el contraste de hipótesis de que todos los coeficientes de regresión logística son cero, mediante un estadístico chi-cuadrado. El estadístico de chi-cuadrado, derivado del cociente de verosimilitudes y de utilidad para comprobar la bondad de ajuste del modelo (la cercanía de los valores predichos por el modelo a los observados) fue 8.64 (p = 0.0344); por consiguiente no se rechaza la hipótesis nula de ue el modelo es acertado. Se calcularon los valores $_{ij}$ (estimación de la probabilidad de que se produzca la primera respuesta, que no se dé la existencia de embriones) en los niveles de las correspondientes variables explicativas. También se contrastó la hipótesis de si los distintos coeficientes de regresión logística son significativos o no ($H_0$: $_j$=0); SAS utiliza $z^2$, el estadístico de Wald, que sigue una distribución Chi-cuadrado con un grado de libertad, al igual que en el estudio de caso uno.

## CONCLUSIONES Y RECOMENDACIONES

La mejor manera estadística de analizar datos de tipo categórico es la regresión logística; de manera tradicional éste tipo de datos se analizan mediante tablas de contingencia, observando en ellas las probabilidades marginales, conjuntas y condicionales, comparando puntualmente las diferencias en ellas. Mientras que con una prueba de independencia con el estadístico chi-cuadrado se puede determinar si las variables explicatorias se relacionan o no con la variable respuesta, concluyendo únicamente si las variables se relacionan o no, más sin medir un grado de relación entre ellas.

Se recomienda llevar el análisis estadístico más allá de eso y utilizar herramientas de regresión logística para poder modelar las probabilidades de interés según sea el caso mediante los estimadores correspondientes a cada modelo mediante la utilización de los modelos log-lineales; con la utilización de los procedimientos CATMOD y LOGISTIC del paquete estadístico SAS es más fácil y rápido realizar los cálculos correspondientes y con ellos poder hacer inferencias más completas a cerca de los datos.

## LITERATURA CITADA

Agresti, A. 2002. Categorical data analysis. Editorial Wiley. New York.

Ato, G. y G. López. 1996. Análisis estadístico para datos categóricos. Editorial Síntesis, S.A. Madrid.

Bishop, Y. M. M., S. E. Fienberg y P. W. Holland. 1975. Discrete multivariate analysis. Cambridge, MA: MIT Press.

Cañón J., 1986. Caracteres discretos en mejora genética animal. Investigación agraria, Produciion y Sanidad Animales 205-236.

Cavestany, D., C. S. Galina y C. Viñoles. 2001. Efecto de las caracteristicas del reinicio de la actividad ovárica posparto en la eficiencia reproductiva de vacas Holstein en pastoreo. Arch Med Ver 32: 21-33.

Contreras B, D. 2014. Comparación de métodos estadísticos en el análisis de datos binarios dentro de programas para la superovulación de tres razas bovinas. Programa especial de investigación. Facultad de Zootecnia y Ecologia. Universidad Autónoma de Chihuahua. México.

Fienberg, S. E. y K. Larntz. 1976. Loglinear representation for paired and multiple comparison models. Biometrika 63: 245-254.

Goodman, L. A. 1970. The multivariate analysis of qualitative data: Interaction among multiple classifications. J. Amer. Statist. Assoc. 65:226-256.

Grizzle, J. E., C. F. Starmer y G. G. Koch. 1969. Analysis of categorical data by linear models. Biometrics 25:489-504.

Hines, W. W. y D. C. Montgomery. 2004. Probabilidad y estadistica para ingenieria. 3ª edición. Editorial CECSA. México

Infante G., S. y G. Zárate. 2000. Métodos Estadísticos. Primera Edición. Editorial Trillas. México.

Jaramillo, G. 2003. Inseminación artificial intrauterina con semen fresco en ovejas primíparas (F1) Dorper-pelibuey. Tesis de Maestría en Ciencias. Universidad Autónoma Chapingo.

Ku, H. H., R. N. Varner y S. Kullback. 1971. Analysis of multidimensional contingency tables. J. Amer. Statist. Assoc. 66:55-64.

Le, C. T. 2003. Introductory Biostatistics. Wiley-Interscience. USA.

Lipschutz, A. y Lipson, A. 2001. Probabilidad. 2ª Edición. Editorial McGraw-Hill. México.

McCullagh, P. y J. A. Nelder. 1983. Generalized linear models. Chapman & Hall. London.

Sahagun, C. J. 1994. Estadística descriptiva y probabilidad: una perspectiva biológica. Universidad Autónoma Chapingo. México

SAS, 2 1. User's Guide. SAS Institute Inc. Cary North carolina. USA.

Silva, B., y J. Cañón. 2000. Análisis de variables categóricas mediante el procedimiento CATMOD de SAS: aplicación a datos de cruzamiento industrial en bovino. Reporte tecnico. Dpto. Producción Animal; Facultad de Veterinaria; Universidad Complutense de Madrid. España.

Stokes, M. E. , C. S. Davis y G. g. Koch. 2000. Categorical Data Analysis Using the SAS System.SAS Institute.

Verde, O. 2000. Comparación de métodos para análisis de datos binomiales en producción animal. Zoo. Trop. 18:3-28.

# APÉNDICE

A1. Entrada de datos en el paquete estadístico SAS del estudio de caso uno

Sintaxis del análisis en el estudio de caso uno

```
data dilut;
input macho dilutor gest count;
cards;
1 1 1 17
1 1 0 3
1 2 1 15
1 2 0 5
2 1 1 16
2 1 0 4
2 2 1 16
2 2 0 4
;
proc logistic descending;
weight count;
class dilutor macho;
model gest=dilutor macho;
output out=tarea1 predprobs=individual;
proc print data=tarea1;
run;
proc catmod;
model gest=macho dilutor;
weight count;
run;
```

A2. Análisis de varianza para el nivel uno de la variable macho

| Estadístico | DF | Valor | Probabilidad |
|---|---|---|---|
| Chi-cuadrado | 1 | 0.625 | 0.43 |
| Ratio chi-cuadrado de la verosimilitud | 1 | 0.630 | 0.43 |
| Adj. chi-cuadrado de continuidad | 1 | 0.156 | 0.69 |
| Chi-cuadrado Mantel-Haenszel | 1 | 0.609 | 0.43 |
| Coeficiente Phi | | -0.125 | |
| Coeficiente de contingencia | | 0.125 | |
| V de Cramer | | -0.125 | |

A3. Estimadores de riesgo relativo (fila1/fila2)

| Tipo de estudio | Valor | 95% Límites de confianza | |
|---|---|---|---|
| Case-Control (Odds Ratio) | 0.5294 | 0.1079 | 2.5983 |
| Cohort (Col1 Risk) | 0.6000 | 0.1651 | 2.1801 |
| Cohort (Col2 Risk) | 1.1333 | 0.8288 | 1.5497 |

A4. Análisis de varianza para el nivel dos de la variable macho

| Estadístico | DF | Valor | Probabilidad |
|---|---|---|---|
| Chi-cuadrado | 1 | 0.00 | 1.0 |
| Ratio chi-cuadrado de la verosimilitud | 1 | 0.00 | 1.0 |
| Adj. chi-cuadrado de continuidad | 1 | 0.00 | 1.0 |
| Chi-cuadrado Mantel-Haenszel | 1 | 0.00 | 1.0 |
| Coeficiente Phi | | 0.00 | |
| Coeficiente de contingencia | | 0.00 | |
| V de Cramer | | 0.00 | |

A5. Entrada de datos en el paquete estadístico SAS, del estudio de caso dos

Sintaxis del análisis en el estudio de caso dos

```
data raza;
input raza protocolo embriones count;
cards;
1 1 0 110
1 1 1 100
1 2 0 189
1 2 1 232
2 1 0 3
2 1 1 3
2 2 0 99
2 2 1 88
3 1 0 12
3 1 1 24
3 2 0 12
3 2 1 20
;
proc freq;
weight count;
tables raza*protocolo*embriones / chisq relrisk;
tables protocolo*raza*embriones / chisq relrisk;
exact pchi or;
run;
proc logistic descending;
weight count;
class raza protocolo;
model embriones= raza protocolo;
output out=tarea1 predprobs=individual;
proc print data=tarea1;
run;
proc catmod;
model embriones= raza protocolo;
weight count;
run;
```

A6. Análisis de varianza para el nivel uno de la variable raza

| Estadístico | DF | Valor | Probabilidad |
|---|---|---|---|
| Chi-cuadrado | 1 | 3.15 | 1.0 |
| Ratio chi-cuadrado de la verosimilitud | 1 | 3.14 | 1.0 |
| Adj. chi-cuadrado de continuidad | 1 | 2.85 | 1.0 |
| Chi-cuadrado Mantel-Haenszel | 1 | 3.15 | 1.0 |
| Coeficiente Phi | | 0.07 | |
| Coeficiente de contingencia | | 0.07 | |
| V de Cramer | | 0.07 | |

A7. Estimadores de riesgo relativo para la raza uno

| Tipo de estudio | Valor | 95% Límites de confianza | |
|---|---|---|---|
| Case-Control (Odds Ratio) | 1.35 | 0.968 | 1.88 |
| Cohort (Col1 Risk) | 1.16 | 0.987 | 1.37 |
| Cohort (Col2 Risk) | 0.86 | 0.732 | 1.02 |

A8. Análisis de varianza para el nivel dos de la variable raza

| Estadístico | DF | Valor | Probabilidad |
|---|---|---|---|
| Chi-cuadrado | 1 | 0.02 | 0.88 |
| Ratio chi-cuadrado de la verosimilitud | 1 | 0.02 | 0.88 |
| Adj. chi-cuadrado de continuidad | 1 | 0.00 | 1.0 |
| Chi-cuadrado Mantel-Haenszel | 1 | 0.02 | 0.88 |
| Coeficiente Phi | | -0.01 | |
| Coeficiente de contingencia | | 0.01 | |
| V de Cramer | | -0.01 | |

A9. Estimadores de riesgo relativo para la raza dos

| Tipo de estudio | Valor | 95% Límites de confianza | |
|---|---|---|---|
| Case-Control (Odds Ratio) | 0.88 | 0.17 | 4.52 |
| Cohort (Col1 Risk) | 0.94 | 0.41 | 2.12 |
| Cohort (Col2 Risk) | 1.06 | 0.47 | 2.39 |

A10. Análisis de varianza para el nivel tres de la variable raza

| Estadístico | DF | Valor | Probabilidad |
|---|---|---|---|
| Chi-cuadrado | 1 | 0.12 | 0.88 |
| Ratio chi-cuadrado de la verosimilitud | 1 | 0.13 | 0.88 |
| Adj. chi-cuadrado de continuidad | 1 | 0.01 | 1.00 |
| Chi-cuadrado Mantel-Haenszel | 1 | 0.12 | 0.88 |
| Coeficiente Phi | | -0.04 | |
| Coeficiente de contingencia | | 0.04 | |
| V de Cramer | | -0.04 | |

A11. Estimadores de riesgo relativo para la raza tres

| Tipo de estudio | Valor | 95% Límites de confianza | |
|---|---|---|---|
| Case-Control (Odds Ratio) | 0.83 | 0.30 | 2.25 |
| Cohort (Col1 Risk) | 0.88 | 0.46 | 1.69 |
| Cohort (Col2 Risk) | 1.06 | 0.74 | 1.52 |