# UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

## ANALISIS ESTADISTICO DE LOS FACTORES GENETICOS QUE INTERVIENEN EN EL DESARROLLO DE LA RETINOPATIA DIABETICA

# T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I A

P R E S E N T A :

**SANDRA ROMERO HIDALGO**

DIRECTOR DE TESIS: M. EN C. JOSE ANTONIO FLORES DIAZ.

VNIVERSIDAD NACIONAL
AVPNOMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

**Comunicamos a usted que hemos revisado el trabajo de Tesis:**
Análisis estadístico de los factores genéticos que intervienen en el desarrollo de la retinopatía diabética

**realizado por** Sandra Romero Hidalgo

**con número de cuenta** 9150551-8 **, pasante de la carrera de** Actuaría

**Dicho trabajo cuenta con nuestro voto aprobatorio.**

**Atentamente**

**Director de Tesis**
**Propietario** M. en C. José Antonio Flores Díaz

**Propietario** M. en A.P. María del Pilar Alonso Reyes

**Propietario** Dr. Julio Granados Arreola

**Suplente** Act. María Guadalupe Tzintzún Cervantes

**Suplente** Act. Tania Chávez Razo

**Consejo Departamental de Matemáticas**

M.en A.P. María del Pilar Alonso Reyes

CONSEJO DEPARTAMENTAL
DE
MATEMÁTICAS

A ti *Elvira*,
que siempre has estado,
que nunca me has fallado,
te digo:
este logro es tuyo.

# AGRADECIMIENTOS

*"Corregir el presente, lo veo complicado, pero para corregir el futuro, hay que empezar desde hoy"*, en este sentido mi primer agradecimiento va dirigido a la Universidad Nacional Autónoma de México, quien tiene una gran responsabilidad, la formación de aquellos que serán el futuro de nuestro país. En lo particular siento un gran orgullo pertenecer a ella y haré todo lo que esté en mi, para siempre con mi ejemplo poner en alto su nombre.

Quisiera agradecer también de manera especial a todos y cada uno de los miembros del jurado, ya que su colaboración y comentarios ayudaron a enriquecer el contenido de este trabajo, en especial a José Antonio Flores, María del Pilar Alonso y Julio Granados.

Hay tres personas que también participaron de manera activa conmigo y que mediante estas líneas quisiera que estuvieran enteradas del gran apoyo que fueron para mi: Franz Turczynski, Lorena Romero y Hugo Quiróz.

En fin, agradezco profundamente a mi familia, maestros y amigos, el apoyo que de manera directa o indirecta me brindaron para la realización de este trabajo y por ende la conclusión de un aspecto realmente importante de mi vida.

# TABLA DE CONTENIDOS

## CAPÍTULO 3



## CAPÍTULO 4

# INTRODUCCIÓN

La Diabetes Mellitus se ubicó en 1990 como la cuarta causa de mortalidad en México, por razones como ésta estudiar las causas de esta enfermedad es un tema importante que durante mucho tiempo se ha venido investigando, no sólo en nuestro país sino en el mundo entero, ya que se ha visto que los factores causales varían de región en región, según el clima, los hábitos alimenticios, se habla también de que depende de factores genéticos, en fin, son muchas las causas que provocan una mayor susceptibilidad al desarrollo de esta enfermedad.

Sin embargo, hay otro tema de igual importancia del cual se sabe muy poco, y que es una de las consecuencias de la diabetes mellitus, la retinopatía diabética; ésta es la principal causa de ceguera en México y, hasta donde se sabe, en la mayor parte de los países occidentales ocupa ese lugar. Los estudios realizados demuestran que cuanto mayor es el tiempo de duración de la diabetes, mayor es la probabilidad de presentar algún grado de retinopatía, sin embargo, no se conoce con exactitud el tiempo que transcurre entre el inicio de la primera y la aparición de la segunda. La importancia radica en que la retinopatía no es reversible pero si controlable, por lo que si se logra detectar en etapas tempranas, la pérdida de la visión no será absoluta.

En la primera parte de este trabajo se presentan antecedentes de las enfermedades mencionadas, así como algunos conceptos de genética; esto con la finalidad de aclarar de antemano algunas de las preguntas que se haría una persona que no tiene ningún conocimiento teórico de la enfermedad.

Temas como estos existen miles en el campo de la medicina, lo que hace resaltar la importancia de la investigación dentro de este campo, la cual a veces se ve limitada por factores económicos, tecnológicos, entre otros. Por lo anterior, el segundo capítulo de este trabajo está dedicado al proceso de investigación, el cual de manera general es el mismo para todas las disciplinas, sin embargo, existen en cada una de ellas condiciones propias de la materia, por ejemplo, en medicina existen principalmente ocho tipos de estudios diferentes, los cuales resultan de analizar ciertas condiciones, a éstas se les llama criterios de clasificación y se encuentra explicados también en dicho capítulo.

Como el título lo indica el objetivo general de esta tesis es que mediante la utilización de algunas técnicas estadísticas se pueda determinar si existen factores genéticos involucrados en el desarrollo de la retinopatía diabética, para tal efecto se tomó el código genético de tres grupos de individuos (personas que no presentaban la enfermedad, personas que presentaban diabetes mellitus pero no retinopatía diabética y personas que presentaban diabetes mellitus y también retinopatía diabética), la manera de obtener los datos, las estructura de éstos, las variables involucradas, así como, la hipótesis sugerida, son temas tratados en el capítulo tres.

Los métodos utilizados para decidir si la hipótesis es correcta o no fueron: *la prueba ji-cuadrada para el análisis de tablas de contingencia con dos criterios de*

*clasificación*, con éste se desea determinar si existe alguna relación entre dos rasgos diferentes en los que una población ha sido clasificada y en donde cada rasgo se encuentra subdividido en cierto número de categorías. El análisis de una tabla de este tipo supone que las dos clasificaciones son independientes, esto es, bajo la hipótesis nula de independencia se desea saber si existe una diferencia suficiente entre las frecuencias que se observan y las correspondientes frecuencias que se esperan, tal que la hipótesis nula se rechace. La prueba ji-cuadrada proporciona los medios apropiados para analizar este tipo de tablas. El segundo método utilizado fue el de *regresión logística* en particular el caso multinomial, el cual contempla una variable respuesta nominal $Y$ que tiene tres categorías (los tres grupos de personas mencionadas anteriormente) y el vector $X$ donde $X = (x_1, x_2, \ldots\ldots x_n)$ para n variables independientes (genes), con este método lo que se obtiene es un modelo mediante el cual se describe el tipo de asociación que existe entre las variables, así como, las variables independientes que están determinando las posibles respuestas. El modelo proporciona también, con base en los datos observados, la probabilidad predictiva de obtener alguna de las posibles respuestas.

Tomando en cuenta el objetivo buscado en este estudio, se consideró que estos dos métodos eran adecuados, una explicación detallada de ellos se presenta también en el capítulo tres.

Por último, los resultados obtenidos de la aplicación de estos métodos se presenta en el capítulo cuatro, así como, los criterios utilizados para seleccionar de todo el conjunto de resultados, los verdaderamente importantes, mismos que se presentan en el apartado de las conclusiones, junto con una explicación del tipo de asociación existente entre el o los genes considerados y la enfermedad.

Para realizar el análisis referente al método de regresión logística, fue necesario utilizar un paquete estadistico (STATA, versión 3.0) del cual se presentan tres corridas en el apéndice 1, las cuales están explicadas en el capítulo cuatro.

# CAPÍTULO 1

## 1.1 Introducción

Para el mejor entendimiento de los capítulos posteriores se dará una breve introducción de genética, así como de la diabetes mellitus, a qué se debe, los tipos de diabetes más frecuentes, los factores de mayor riesgo, sus principales manifestaciones, etc. Entre las manifestaciones más importantes se encuentra la retinopatía diabética, la cual es tema principal de este trabajo, por lo que también se consideran algunos conceptos de esta última y el porque de la importancia de una detección y prevención oportuna.

## 1.1 Genética

Los principios fundamentales sobre la herencia fueron descubiertos por Gregorio Mendel (1822-1884), quien empezó en 1856 usando chícharos para investigar de que manera ciertas características eran heredadas, llamándolas a unas dominantes y a otras recesivas[1]

[1] La dominancia de una caracteristica sobre otra es común pero no es un fenómeno universal.

Mendel introdujo también las siguientes hipótesis, características contrastantes, tales como la redondez o rugosidad de los chicharos, son determinadas por factores llamados genes que son transmitidos de los padres a los hijos. Cada planta de chicharo tiene dos genes para la forma de la semilla, uno heredado por la parte masculina y otro por la femenina, cada uno de estos puede ser de dos tipos o alelos distintos, uno el que determinan las semillas redondas (alelo de redondez), y el otro las rugosas (alelo de rugosidad).

En 1902 Walter S. Sutton en Estados Unidos de Norteamérica y Theodor Boveri en Alemania independientemente uno del otro, sugirieron que los genes están contenidos en los cromosomas. La posición que éstos tienen en el cromosoma es conocida como locus (lugar en latín); los genes son los portadores de la herencia.

Cada especie tiene una enciclopedia integrada por un número específico de tomos o cromosomas. La del caballo consta de 64, la del rinoceronte tiene 82 y los seres humanos poseen 46 cromosomas distribuidos en 23 pares. El patrón genético exclusivo de cada uno de ellos está integrado por determinadas características que comparten con los miembros de su especie y raza, pero también incluye características que lo hacen ser individual, a excepción de los gemelos idéntico o monocigóticos, es decir, que provienen de un sólo cigoto.

Es importante introducir otros dos términos del vocabulario genético. *Homocigoto* es un individuo en el cual los dos genes de un par (un gen heredado de cada padre) son iguales, es decir, un individuo con dos alelos idénticos. *Heterocigoto* es un individuo con dos alelos diferentes.

En 1909 Wilhelm Johannsen introdujo la distinción entre fenotipo y genotipo. El *fenotipo* de un organismo es la apariencia que se puede observar: es morfológica, fisiológica y de comportamiento. El *genotipo* es el código genético que ha sido heredado. Durante la vida de un individuo, el primero puede cambiar; mientras que el segundo permanece constante. Debe tomarse en cuenta que la relación entre estos dos no es fija, esto es porque el primero es el resultado de una compleja red de interacciones entre diferentes genes; en general los individuos no tienen fenotipos idénticos, aunque pueden ser similares cuando sólo se consideran ciertas características.

El sistema mayor de histocompatibilidad, o de modo abreviado, HLA (HLA, del inglés: Human Leukocyte Antigens) es un complejo de genes cuyos sitios específicos están localizados en el brazo corto del cromosoma 6. En la cartografía del sistema HLA pueden definirse 4 subsistemas génicos o clases de productos de diferente estructura y variada función en el control y modulación de la respuesta inmune. Para este estudio en específico se tomaron en cuenta los genes (loci) de la Clase I denominados A, B, C y los de la Clase II DQ, DR.

Conviene destacar que los loci están situados físicamente muy próximos, lo que ocasiona que estos genes estén vinculados. Pero lo notable en el sistema HLA es el llamado fenómeno de desequilibrio de ligamiento, que puede definirse como el hecho de que ciertas parejas de alelos o antígenos[2] HLA se expresan juntos con mayor frecuencia de la que podría esperarse por azar, tanto en familias de individuos como en poblaciones. Una interpretación genérica de este fenómeno, es que fuerzas selectivas a lo largo de la evolución tienden a evitar la recombinación de

[2] Sustancia que introducida en un organismo, puede provocar la formación de anticuerpos.

ciertas variantes alélicas en el sistema HLA con otros loci a pesar de la distancia entre ellos o dicho de otro modo, la selección tiende a retener sólo a aquellas combinaciones de antígenos que son potencialmente ventajosas para la defensa inmunológica.

## 1.2 Diabetes Mellitus

La diabetes mellitus, las enfermedades del corazón, los tumores malignos y los accidentes, entre otras, son las causas de mayor número de muertes en México. La magnitud del problema hace resaltar la importancia que tiene en primer nivel la educación, prevención, diagnóstico, tratamiento y rehabilitación del paciente en diversas enfermedades, pero sobre todo en la que ocupa este trabajo, la diabetes mellitus.

La diabetes mellitus es un síndrome[3] multiforme. Puede aparecer a edades tempranas o avanzadas de la vida, ser resultado de un proceso autoinmunitario asociado a una predisposición genética y desencadenado por factores ambientales hasta ahora desconocidos o bien, puede obedecer a una disminución en la sensibilidad a la acción de la insulina. Entre las manifestaciones crónicas más importantes están las que resultan del daño de los nervios (neuropatía), de los vasos sanguíneos pequeños (microangiopatía), tanto en riñones (nefropatía) como retina (**retinopatía**) entre otras.

[3] Conjunto de síntomas de una enfermedad

Existen tanto formas primarias como secundarias de la diabetes mellitus. Las primeras son aquellas no relacionadas con otra condición que las cause o las modifique, es decir, son el resultado de una susceptibilidad individual determinada genéticamente para expresar la enfermedad. Estas formas primarias corresponden a las subclases:

1. Diabetes mellitus insulinodependiente (DMID) o tipo I.
2. Diabetes mellitus no insulinodependiente (DMNID) o tipo II.

### 1.2.1 Diabetes mellitus insulinodependiente (DMID)

Denominada diabetes juvenil, es la forma más grave de la diabetes mellitus primaria y la menos frecuente, en México constituye alrededor del 1 al 2% de todos los casos de diabetes. Suele diagnosticarse en la infancia o en la adolescencia, por lo que tiene un fuerte impacto en la dinámica familiar y más tarde en la capacidad productiva del individuo.

Algunos de los factores de mayor riesgo para tener DMID son:

1. Ser gemelo monocigoto de un individuo con DMID.
2. Ser hermano de un individuo con DMID y ser:
   a) HLA idéntico (que comparte ambos haplotipos).
   b) Haploidéntico (que comparte un haplotipo).
   c) HLA no idéntico (que no comparte haplotipos).
3. Ser hijo de un individuo con DMID.

Factores genéticos y ambientales. El riesgo de desarrollar DMID está relacionado con ciertos genes de la clase II del complejo mayor de histocompatibilidad. Se ha observado que la asociación de DMID con genes específicos del locus HLA-DR es probablemente secundaria al fenómeno de desequilibrio de ligamiento con genes del locus HLA-DQ. La fuerza de asociación con un haplotipo específico no es la misma en distintos grupos étnicos y disminuye con la edad de presentación de la DMID. Existen también factores distintos de los genéticos, ya que únicamente la mitad de los gemelos monocigotos de pacientes con DMID desarrollan la enfermedad, ciertas enfermedades virales se han identificado también como probables desencadenantes del proceso.

Etiología. Es multifactorial, resultante de la interacción entre factores ambientales probables, como virus, tóxicos, dietas, o supuestos, así como una predisposición genética individual que, en mayor o menor grado, predomina en distintos grupos étnicos y condiciones geográficas, por ejemplo, existe una mayor vulnerabilidad e incidencia en los países más distantes del Ecuador (Finlandia, Noruega) y una cierta indemnidad[4] en las áreas más próximas a aquél (Cuba, Perú), con esto se puede decir que existe una incidencia decreciente de norte a sur. Este fenómeno no es constante, pero si frecuente.

Es predominante en los grupos de edad entre 0 y 29 años y dentro de este margen existe una vulnerabilidad en las poblaciones infantil y adolescente, esto es

[4] Seguridad que se da a alguien, de que no sufrirá daño.

en cualquier circunstancia geográfica y cualquier grupo étnico. También es un síndrome de otoño, de invierno, o de ambas estaciones incluyendo a Cuba donde la distinción en las características climáticas es mucho menos nítida que en otras partes del mundo.

### 1.2.2 Diabetes mellitus no insulinodependiente (DMNID)

Antes denominada diabetes del adulto. Esta es la forma más frecuente de diabetes mellitus (98 a 99% del total, en México). Suele iniciarse después de la cuarta decada de vida y su prevalencia aumenta con la edad. La mayoría (80 a 85%) de los pacientes con DMNID son obesos en el momento del diagnóstico y una minoría está en su peso ideal.

En México la prevalencia[5] en adultos de todas las edades es de 8 a 10%, pero uno de cada cuatro individuos mayores de 50 años tiene diabetes, con cierta predominancia del sexo femenino. Preocupa también la elevada prevalencia de diabetes (5%) en individuos relativamente jóvenes (35 a 45 años). Estas cifras son aún mayores en la población mexicana que emigró a los Estados Unidos, donde la prevalencia de diabetes prácticamente se ha duplicado, lo cual quizá se relaciona con los cambios de hábitos de vida, en particular los alimenticios y de ejercicio, que favorecen un incremento en la masa corporal.

[5] La prevalencia de una enfermedad es el número de casos que hay en un punto determinado del tiempo. La tasa de prevalencia es el cociente del número de casos entre el número de personas estudiadas inicialmente.

Pese a que la diabetes es un problema común, un gran porcentaje de las personas que la padecen (alrededor del 30 al 40%) no han sido diagnosticadas, y para detectarlas se requieren estudios de escrutinio[6].

Algunos factores de riesgo para desarrollar DMNID son:

1. Ser gemelo monocigoto de un individuo con DMNID.
2. Ser familiar en primer grado (hermano, padre o hijo) de un individuo con DMNID.
3. Ser obeso.
4. Ser madre de un producto que pesó al nacer 4.5 Kg. o más.
5. Ser miembro de un grupo étnico con alta prevalencia de DMNID.

Factores genéticos y ambientales. La susceptibilidad para desarrollar DMNID tiene un evidente componente hereditario. La enfermedad ocurre más a menudo en los familiares de un individuo afectado que en la población general. La frecuencia de concordancia de DMNID en gemelos monocigotos es de por lo menos 70% y en algunas series alcanza casi el 100%. A pesar de esto no se ha podido identificar un patrón mendeliano definido de transmisión. Los principales factores adquiridos que contribuyen al desarrollo de DMNID son obesidad, inactividad física, embarazo y edad avanzada.

[6] Examen o averiguación de una cosa

## 1.3 Retinopatía Diabética

Una de las consecuencias de la diabetes mellitus es la retinopatía diabética, la cual es la principal causa de ceguera en México y, hasta donde se sabe, lo mismo ocurre en la mayor parte de los países del mundo occidental.

Las investigaciones sugieren que existen dos razones importantes que explican este hecho. La primera es un problema de educación y salud pública: individuos con diabetes mellitus de larga evolución dejan de dar importancia a sus síntomas y tienden a acudir con menos frecuencia a consulta médica. La segunda está ligada a la naturaleza de la enfermedad crónica y progresiva: el inicio de cambios oftalmológicos no tiene prioridad en el diagnóstico del paciente diabético.

Esta complicación es la más importante de aquellas que afectan al ojo, es la causante de la ceguera de tipo irreversible. No se conoce con exactitud el tiempo que media entre el inicio de la diabetes y la aparición de la retinopatía, por lo que ésta es impredecible. Lo que se sabe es que cuanto mayor es el tiempo de duración de la diabetes, más grandes son las posibilidades de presentar algún grado de retinopatía.

La retinopatía puede ser proliferativa o no proliferativa. La primera se encuentra más frecuentemente en los DIMD y es la que produce pérdida grave de la visión. De aquí que la detección, el estudio, la comprensión, el seguimiento y la atención correcta de los pacientes con la forma no proliferativa son en la actualidad los aspectos de mayor importancia en el abordaje oftalmológico del paciente diabético.

En pacientes a quienes se les diagnostica la diabetes antes de los 30 años de edad, sin considerar que sean insulinodependientes, la incidencia[7] de retinopatía diabética después de cuatro años de evolución es de aproximadamente 60%, después de los 10 años es de 70% y después de los 20 años de 97% aproximadamente. De pacientes a quienes se les diagnostica retinopatía diabética no proliferativa en su examen inicial, después de cuatro años, el 55% permanece sin cambios, el 40% progresa moderadamente y el 11% desarrolla la forma proliferativa. En pacientes en quienes se detecta retinopatía diabética proliferativa en su examen inicial, el 14% muestra signos de alto riesgo para la pérdida grave de la visión.

La incidencia de esta enfermedad en quienes la diabetes se diagnosticó después de los 30 años de edad varía, entre aquellos que requieren insulina y los que no la necesitan. A los cuatro años de evolución, para aquellos que usan insulina y no muestran retinopatía diabética en un examen oftalmológico inicial, se ha detectado que alrededor del 47% la ha desarrollado, en 34% progresa de manera moderada y en el 7% se desarrolla la forma proliferativa. En pacientes no dependientes de insulina, en 34% se desarrolla la retinopatía, en 25% progresa moderadamente y en 2% se presentan cambios proliferativos.

[7] Es el número de casos nuevos de un evento que ocurren durante un determinado periodo de tiempo

Se cuenta con información que señala que en las últimas décadas el número de pacientes diabéticos no insulinodependientes se incrementó en forma significativa en México, lo cual repercute en todas las instituciones de salud al constituir una de las causas principales de consulta médica y de admisión hospitalaria. Tal vez el mexicano tiene una mayor predisposición genética para el desarrollo de la enfermedad y el cambio de sus hábitos de vida, caracterizados por un mayor sedentarismo[8] y sobrepeso, se acompaña de una mayor prevalencia de diabetes, tal y como ocurre en las grandes ciudades mexicanas y en la población nacional que emigró al sur de Estados Unidos.

[8] Sedentario. De poca agitación o movimiento.

# CAPÍTULO 2

## 2.1 Introducción

Este capítulo tiene como objetivo recordar en primera instancia el famoso método científico, con la finalidad de hacer notar la importancia de todos y cada de los pasos de éste, y en específico de la importancia que tiene en una investigación la parte estadística, que a veces se olvida. Este método científico es de manera general el procedimiento seguido en todos los proyectos de investigación, sin embargo, dependiendo de la materia de que se trate se toman criterios específicos. El campo de interés en este trabajo es la medicina, por lo que, se mencionarán también los criterios considerados en este campo, *determinando cuales de éstos son válidos para el estudio analizado más adelante.* Por otro lado, se mencionarán también los métodos estadísticos más comúnmente utilizados.

## 2.2 El método científico

El objetivo de la investigación es descubrir respuestas a determinadas interrogantes a través de la aplicación de procedimientos científicos, los cuales siguen de manera general los pasos del método científico, sin embargo, por la variedad de temas que requieren de investigación, todos con particularidades específicas han surgido variaciones a este método, sin negar su importancia en el mundo de la ciencia.

Conceptualmente, el método científico no es otra cosa que el conjunto de normas y principios generales que deben seguirse en una investigación. El paso número uno de este método es la **observación empírica** que puede ir desde algo sencillo como que la madera flota, hasta observaciones verdaderamente complejas. Sin embargo, la continuidad del hombre depende en gran medida de estas observaciones y de las consecuencias que ha traido haberlas observado. El punto importante resulta cuando derivado de la observación de un hecho o serie de hechos, cabe preguntar si los hechos de determinada clase siguen siempre el mismo modelo, o si bien existen circunstancias en las que el resultado puede ser diferente.

Las razones para efectuar preguntas que lleven a la investigación son de dos clases: razones intelectuales, basadas en el deseo de saber o entender por la pura satisfacción del conocimiento o comprensión; y razones prácticas, fundadas en el deseo de saber para ser capaces de hacer mejor o de forma más eficaz alguna cosa. Las investigaciones que llevan a estos tipos de cuestiones son:

investigación pura o básica y aplicada, la empresa científica ha tenido en cuenta a la primera por su valor intrínseco[9] y a la segunda por cuanto contribuye a obtener beneficios prácticos.

El segundo paso en este método es la **formulación de hipótesis**, ésta como definición es una suposición que permite establecer relaciones entre los hechos. Las hipótesis parten de suposiciones o conjeturas más o menos casuales que se denominan hipótesis previas. Éstas se van fundamentando en observaciones científicas, a través de las cuales se llega a identificar claramente a los elementos que intervienen para que un fenómeno se presente. Las hipótesis generales resultan de observaciones rigurosas y comprobaciones realizadas no sólo en grupos relativamente grandes sino sobre una población en general.

En todos los casos la hipótesis implica una verdad provisional y tiene como base la experiencia puesto que surge y se fundamenta en la observación empírica. La ausencia de contradicción es un requisito fundamental, existen dos tipos de contradicciones, uno es la interna, es decir, la hipótesis da lugar a dos conclusiones incompatibles, o la externa, donde la hipótesis es incompatible con los hechos.

Otro punto importante es la **identificación y control de las variables**. Es preciso identificar y definir con precisión las variables, pues de lo contrario la investigación se verá contaminada por factores extraños, y conducirá a conclusiones falsas.

[9] Es el valor que tiene por si mismo.

La variable independiente es la variación o alteración que el investigador introduce en su experimento. El control de esta variable es una característica esencial, en ello radica la diferencia entre el método descriptivo y el método experimental. La variable dependiente es la consecuencia de las alteraciones en la o las variables independientes.

Existen también variables que actúan adicionalmente a la variable independiente y que pueden afectar a la dependiente, éstas son: las variables externas o de control, la variable interventora y la variable moderadora. Deben eliminarse las variables externas para detectar claramente la influencia de la variable independiente sobre la dependiente. Las variables interventoras no se pueden medir en forma directa, sino a través de los efectos que tiene sobre el fenómeno (aprendizaje, motivación, etc.). La variable moderadora o variable dependiente secundaria puede alterar el efecto de la variable independiente sobre la dependiente, aunque no es la responsable directa (factores nutricionales, sexo, raza, etc.).

Además de las variables arriba mencionadas existen otros factores perturbadores como son el paso del tiempo, hechos contingentes, selección no aleatoria de los sujetos, alteraciones en los instrumentos o normas de medición.

El paso siguiente se refiere a la **comprobación de las hipótesis**, la función de la hipótesis es la de afirmar una relación determinada entre fenómenos, de tal modo que ésta se pueda poner a prueba y así concluir que dicha proposición es correcta o incorrecta. Esta comprobación se puede realizar de dos formas, puede ser de manera empírica o de manera numérica.

El método fundamental para realizar la comprobación empírica es planteando la investigación de tal forma que la lógica exija la aceptación o el rechazo de la hipótesis. Existen dos métodos principales para llevar a cabo esta comprobación, el primero es el método del consenso que dice de manera general lo siguiente, si en todos los casos en que se encuentra la condición C se puede hacer la observación Z, cabe llegar a la conclusión de que todos ellos están relacionados causalmente, de igual manera existe el método del consenso negativo, que afirma que cuando se encuentra que la condición de no-existencia de C va asociada a la observación de no-existencia de Z, se puede afirmar una relación causal entre C y Z. El segundo es el llamado método de la diferencia, con esto se afirma que si hay dos o más casos, y en uno de ellos se puede hacer la observación Z, mientras que no es posible hacerla en las demás, y si el factor C se halla presente cuando se hace la observación Z, y no cuando no se hace esta misma observación, entonces se puede afirmar que existe una relación causal entre C y Z. Sin embargo, estos métodos no funcionan si el estudio en cuestión es un poco más complejo o con muchas variables.

La otra forma de realizar la comprobación de la hipótesis es haciendo uso de la estadística, el aspecto más importante de ésta es la obtención de conclusiones basadas en datos experimentales, este proceso se le conoce como inferencia estadística. En estadística la inferencia es inductiva porque se proyecta de lo específico (muestra) a lo general (población). En un procedimiento de esta naturaleza siempre existe la posibilidad de error, nunca podrá tenerse el 100% de seguridad sobre una proposición que se base en la inferencia estadística. Sin embargo, lo que hace que la estadística sea una ciencia es que, unida a cualquier proposición, existe una medida de confiabilidad de ésta. Por último el producto de la investigación científica es el conocimiento científico, éste se expresa a través de **leyes, modelos y teorías**.

Una ley es una afirmación que expresa una relación constante entre algunas variables, entendiendo por ello que existe una conexión forzosa entre ellas.

Los modelos son estructuras (materiales o conceptuales) que presentan similitudes con las características importantes del objeto o proceso que se estudia Un modelo es la expresión simbólica o material de una situación o proceso determinado. Esta expresión no debe ser análoga a la situación o proceso dados, sino más bien debe traducir los hechos observables en un plano distinto, que pueda ser manipulable, seleccionando aquellos aspectos esenciales y representándolos de manera simplificada, a fin de que permita explicar y predecir los fenómenos.

Las teorías son el producto más acabado del conocimiento científico, pues satisfacen integralmente las condiciones de la ciencia. Las teorías sistematizan el conocimiento racional y unificado, haciendo posible la explicación y predicción de fenómenos reales. Es frecuente que los investigadores formulen hipótesis y realicen comprobaciones o refutaciones a través de leyes, modelos y teorías establecidas, debido a que se trata de estructuras de diversos niveles de abstracción que organizan los datos conocidos en un campo determinado aunque no existe evidencia de las relaciones entre las entidades que las conforman.

La función principal de los modelos es facilitar la comprensión de la teoría. El modelo representa la hipótesis acerca de los hechos y las relaciones que se dan entre éstos. El tipo de procedimiento empleado para la construcción de un modelo difiere de una área de conocimiento a otra.

Estos pasos del método científico se siguieron de manera general en el estudio, interese de esta tesis. Los objetivos, las hipótesis propuesta, la definición de las variables, así como, los métodos estadísticos utilizados para la comprobación de dicha hipótesis, se presentan detalladamente en el siguiente capítulo.

## 2.3 Criterios utilizados en el campo de la medicina

En el campo de la medicina se emplean o consideran ciertos criterios de clasificación, que dan lugar a diferentes tipos o protocolos de investigación, éstos son cuatro principalmente y se presentan a continuación:

A. El periodo en que se capta la información,

B. La evolución del fenómeno estudiado,

C. La comparación de poblaciones y

D. La interferencia del investigador en el estudio.

Como en todo proyecto de investigación es muy importante tener siempre en cuenta los objetivos que se pretenden alcanzar, pero también los recursos de los que se dispone, ya que a veces éste es un factor determinante dentro del estudio.

A. De acuerdo con el periodo en que se capta la información, el estudio puede ser:

   a) **Retrospectivo.** Estudio cuya información se obtuvo anteriormente a su planeación con fines ajenos al trabajo de investigación que se pretende realizar.

   b) **Retrospectivo parcial.** Estudio que cuenta con una parte de la información; el resto está por obtenerse.

   c) **Prospectivo**. Estudio en el que toda la información se recogerá, de acuerdo con los criterios del investigador y para los fines específicos de la investigación, después de la planeación de ésta.

B. De acuerdo con la evolución del fenómeno estudiado, el estudio puede ser:

   a) **Longitudinal.** Estudio en que se mide en varias ocasiones la o las variables involucradas. Implica el seguimiento, para estudiar la evolución de cada una de las unidades en el tiempo.

   b) **Transversal.** Estudio en el cual se mide una sola vez la o las variables; se miden las características de uno o más grupos de unidades en un momento dado, sin pretender evaluar la evolución de esas unidades.

C. De acuerdo con la comparación de las poblaciones, el estudio puede ser:

   a) **Descriptivo**. Estudio que sólo cuenta con una población, la cual se pretende describir en función de un grupo de variables y respecto de la cual no existen hipótesis centrales. Quizá se tienen algunas hipótesis que se refieran a la búsqueda sistemática de asociaciones entre varias variables dentro de la misma población.

b) **Comparativo.** Estudio en el cual existen dos o más poblaciones y donde se quieren comparar algunas variables para contrastar una o varias hipótesis centrales. Los estudios comparativos, en lo que toca a la forma de abordar el fenómeno, se dividen en:

De causa a efecto. Se investigan dos o más grupos de unidades de estudio que se diferencian en varias modalidades (p.ej. nada, regular, mucho) de un factor causal y se estudia el desarrollo de éstas para evaluar, conocer y analizar el efecto y la frecuencia de aparición de aquél dentro de cada grupo.

De efecto a causa. Se parte de dos o más grupos de unidades de estudio que presentan cierto fenómeno considerado como efecto en varias modalidades (por ejemplo, presente-ausente) y se retrocede al pasado para determinar o conocer el factor causal, y la proporción en que éste se presentó en los diferentes grupos.

D. De acuerdo con la interferencia del investigador en el fenómeno que se analiza, el estudio puede ser:

a) **Observacional.** Estudio en el cual el investigador sólo puede describir o medir el fenómeno estudiado; por tanto, no puede modificar a voluntad propia ninguno de los factores que intervienen en el proceso.

b) **Experimental.** Estudio en el que el investigador modifica a voluntad una o algunas variables del fenómeno estudiado; generalmente, modifica aquellas consideradas como causa dentro de una relación de causa a efecto. El aspecto fundamental de este tipo de estudio es que se pueden asignar al azar las unidades a las diversas variantes del factor causal.

**Tabla T.1.1** Muestra los diferentes tipos de estudios.

| Características del estudio | | | | Nombre común | P** |
|---|---|---|---|---|---|
| Observacional | Prospectivo | Transversal | Descriptivo | Encuesta Desc. | 1 |
| | Retrospectivo | | Comparativo | Encuesta Comp. | 2 |
| Observacional | Retrospectivo | Longitudinal | Descriptivo | Revisión de casos | 3 |
| Observacional | Retrospectivo | Longitudinal | Comparativo (e-c)* | Casos y controles | 4 |
| Observacional | Retrospectivo | Longitudinal | Comparativo (e-c)* | Perspectiva histórica | 5 |
| Observacional | Prospectivo | Longitudinal | Descriptivo | Estudio de 1 cohorte | 6 |
| | | | Comparativo | Estudio varias cohor. | 7 |
| Experimental | Prospectivo | Longitudinal | Comparativo | Experimento | 8 |

** P = Protocolo, *(e-c) = de efecto a causa.

Como ya se dijo, la combinación de los cuatro criterios anteriores pueden formar los diferentes tipos de protocolos de investigación más comunes, los cuales se muestran en la tabla T.1.1. Analizando ésta y las características del estudio (interés de esta tesis) presentadas en el siguiente capítulo, se puede ubicar éste en el protocolo 2 llamado "Encuesta comparativa prospectiva", es decir, es *observacional*, ya que los datos fueron el resultado del análisis de una muestra de sangre de los pacientes, siendo éstos los que se utilizaron para el análisis, por otro lado, dichos datos fueron obtenidos específicamente para el objetivo del estudio, por lo tanto, es *prospectivo*. Ya que es el código genético lo que se obtiene, las variables son medidas una única vez o no se modifican con el tiempo entonces también es *transversal*, y por último existen tres grupos de estudio, por lo que es *comparativo* (de efecto a causa).

Entre las ventajas sugeridas de este tipo de protocolo están, el que permite formular hipótesis de asociación, es útil para preparar posteriormente un estudio longitudinal comparativo, con la finalidad de contrastar hipótesis; que la representatividad que se obtiene es buena, dependiendo claro de la representatividad de la muestra. Por el lado de las desventajas están, el tiempo, el costo y el no permitir conocer la evolución del fenómeno en estudio.

La selección del método estadístico a utilizar es importante y estará en función de los objetivos del estudio. Los métodos que se utilizan más frecuentemente en este tipo de estudios son los siguientes:

- Prueba $X^2$.
- Prueba exacta de Fisher.
- Medidas como el riesgo relativo[10], riesgo atribuible[10] y razón de momios[10].
- Modelos lineales (logísticos y logarítmicos).

El primer y cuarto punto se encuentran detallados en el siguiente capítulo, ya que fueron utilizados para el análisis, del cual los resultados se presentan en el capítulo cuatro, el segundo punto es un complemento del primero, el criterio es que si los valores esperados son pequeños se utilizará entonces esta prueba. El tercer punto son razones derivadas principalmente de la tasa de incidencia, la cual se refiere al número de casos nuevos que desarrollan la enfermedad durante un intervalo de tiempo.

[10] Méndez Ramírez, Ignacio. *"El protocolo de investigación"*, Ed. Trillas, pag. 163-170.

# CAPÍTULO 3

## 3.1 Objetivo General y Específico

En el primer capítulo se dio una breve introducción de lo que es la retinopatía diabética, así como de la diabetes mellitus, ya que la primera es consecuencia de la segunda y además es la principal causa de ceguera en México, lo anterior con la finalidad de hacer evidente la necesidad y preocupación de médicos y pacientes por conocer mejor los factores que determinan la retinopatía, ya que como se dijo no es reversible únicamente es controlable. También se habló un poco de genética, debido a que el *objetivo general* de este trabajo es encontrar factores genéticos que estén provocando una mayor predisposición al desarrollo de esta enfermedad, tomando para esto diferentes grupos de personas con características específicas, las cuales se mencionan más adelante.

Como se mencionó en el capítulo dos, el análisis estadístico de los datos es una de las partes que integran una investigación y es el *objetivo específico* de esta tesis, es decir, buscar y aplicar métodos estadísticos que se adecuen a los datos

que se obtuvieron para los fines de este estudio, específicamente lo que se busca es poder determinar si la presencia o ausencia de ciertos genes o una combinación de ellos esta asociada con el desarrollo de la retinopatía diabética.

## 3.2 Hipótesis

Derivado de estudios que se han realizado en diferentes países referente a la diabetes y a la retinopatía diabética, se ha visto que los factores determinantes en el desarrollo de estas enfermedades, difieren de una población a otra, la variación es principalmente debido al grupo racial y a la región geográfica. En México por ejemplo, el desarrollo de la retinopatía diabética en pacientes diabéticos al parecer es mayor que en otros países y se tiene la sospecha de que la raza mestiza presenta una mayor predisposición a esta enfermedad. Por consecuencia, se parte de la *hipótesis* de que probablemente existe alguna asociación entre variables, más adelante se explicará cuales son dichas variables, como es que fueron obtenidos los datos, etc. Cabe mencionar que por haberse presentado algunos problemas en el desarrollo del estudio, no se cuenta con una muestra representativa. Sin embargo, los resultados obtenidos aquí serán de ayuda para continuar estudiando y analizando este problema.

## 3.3 Desarrollo del estudio

Primeramente, se seleccionó de los expedientes de la Asociación para Evitar la Ceguera en México[11] a las personas que cumplían con los criterios de inclusión considerados para este estudio, posteriormente se le extrajo una muestra de sangre

[11] La información fue tomada de esta institución debido a que es un hospital de concentración que recibe un número importante de pacientes con retinopatía diabética.

a cada individuo y con base en los resultados de ésta, se definieron las variables involucradas, las categorías de dichas variables y el tamaño real de la muestra. Por último se buscaron los métodos estadísticos que fueran adecuados para el análisis de los datos y con esto poder determinar si la hipótesis sugerida es correcta o no, con base en los resultados que se presentan en el capítulo siguiente.

### 3.3.1 Criterios de inclusión

Se tomaron en cuenta las características siguientes para pertenecer al grupo de estudio,

a) Personas que no presentaran diabetes mellitus y por lo tanto tampoco retinopatía diabética, debido a que como ya se dijo la primera es consecuencia de la segunda (CONTROLES).

b) Personas con muchos años de habérseles diagnosticado la diabetes mellitus y sin presencia de retinopatía diabética (DMSRD).

c) Personas con las características inversas al punto anterior, es decir, personas con pocos años de diabetes y que presentaban mucha retinopatía diabética (DMCRD).

Los dos últimos puntos son debido a que el riesgo de desarrollar retinopatía en pacientes diabéticos se incrementa con el tiempo, por lo que se tiene la idea de que tomando estos criterios de selección sea mayor la evidencia genética que se busca.

### 3.3.2 Variables

De la muestra de sangre extraída a cada elemento del estudio, se obtuvo el código genético registrado en la región mapeada del brazo corto del cromosoma 6, el cual representa cerca del 0.001 del genoma humano, y donde se encuentran los genes controladores de la susceptibilidad a ciertas enfermedades.

Se ha identificado que los genes importantes pertenecen principalmente a las regiones denominadas A, B, C y D, sin embargo, se ha visto que esta última se divide a su vez en subregiones, de las cuales únicamente se considerarán para este estudio las denominadas DR y DQ, por lo que, la información resultante del análisis realizado a la muestra de sangre de cada persona (CONTROLES, DMSRD Y DMCRD) que forma la muestra, se presenta de la siguiente manera:

1: A#, B#, C#, DR#, DQ#

A#, B#, C#, DR#, DQ#

en donde una fila es la información heredada por el padre y la otra por la madre, de esta forma sólo fueron entregados los datos de los CONTROLES, los datos obtenidos de los DMSRD y DMCRD se entregaron incompletos y en forma desordenada donde no se podía determinar cuales provenían del padre y cuales de la madre.

Los alelos considerados para cada una de las variables A, B, C, DR y DQ se presentan a continuación. La asignación de las letras y los números de las variables obedecen al momento de su descubrimiento, mientras que la x corresponde a que no se pudo identificar el alelo de que se trata.

**A** : A1, A2, A3, A9, A10, A11, A19, A28, Ax.

**B** : B5, B7, B8, B12, B13, B14, B15, B16, B17, B18, B21, B22, B27, B35, B37, B40, B41, B48, B70, Bx.

**C** : C1, C2, C3, C4, C5, C6, C7, C8, C10, Cx.

**DR** : DR1, DR2, DR3, DR4, DR5, DR6, DR7, DR8, DR9, DR10, Drx.

**DQ** : DQ1, DQ2, DQ3, DQ4, DQx.

entonces A, B, C, DR y DQ cuentan con 8, 19, 9, 10 y 4 alelos respectivamente, sin considerar dentro de éstas las x, ya que, no es de interés para este estudio saber el comportamiento de esta categoría.

### 3.3.3 Tamaño de la muestra

El tamaño de la muestra se tomó con base en el presupuesto con el que se contaba, debido a que el análisis que se le realiza a la muestra de sangre es muy costoso. Por lo que, se tomaron por un lado 50 CONTROLES, de los cuales se cuenta con la información completa y ordenada, se consideraron también 50 personas con DMSRD y DMCRD, de las cuales sólo se contaba con la información completa de 12 de las 50 personas consideradas, es decir, que tenían las 10

variables de estudio, 6 de los 50 solo contaban con las variables A, B, DR y DQ, y 17 pacientes tenían únicamente DR y DQ aportando únicamente 4 de las 10 variables originales, de las restantes 15 personas no se obtuvo información, ya que los recursos no fueron suficientes.

### 3.3.4 Tipo de investigación

Como ya se mencionó anteriormente, se estudiaron principalmente dos métodos para el análisis de los datos, uno es el análisis de tablas de contingencia, dentro de éste existen tres modalidades. Tablas de contingencia de dos dimensiones, multidimensionales y modelos log-lineales, de estos se utilizó únicamente el primero, las causas se explican más adelante junto con una breve introducción del objetivo de cada uno. El segundo método utilizado fue el de regresión logística, del cual también se introduce una explicación.

#### 3.3.4.1 Análisis de Tablas de Contingencia

El objetivo de éste es poder determinar si existe alguna relación entre dos características diferentes en las que una población ha sido clasificada y en donde cada característica ha sido subdividida en cierto número de categorías. A primera vista este tipo de análisis cumple con el objetivo buscado en el presente estudio, donde las categorías de la primera característica (la presencia o ausencia de la enfermedad) son CONTROLES, DMCRD y DMSRD, y las categorías de la segunda característica son cada uno de los genes considerados. Se dará a continuación una explicación de las tres modalidades de este análisis, a pesar de que únicamente se

utilizó la primera, las otras dos también cumplen con el objetivo, sin embargo, no se utilizaron debido principalmente al tamaño de la muestra, por lo que, si se realizara en un futuro el mismo tipo de estudio pero con mayor número de participantes se podría considerar la posibilidad de utilizarlos.

La forma general que tienen las tablas de dos dimensiones es como sigue:

| variable 1 \ variable 2 | 1 | 2 | . | . | $c$ | *total* |
|---|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | . | . | $n_{1c}$ | $n_{1\bullet}$ |
| 2 | $n_{21}$ | $n_{22}$ | . | . | $n_{2c}$ | $n_{2\bullet}$ |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| $r$ | $n_{r1}$ | $n_{r2}$ | . | . | $n_{rc}$ | $n_{r\bullet}$ |
| *total* | $n_{\bullet 1}$ | $n_{\bullet 2}$ | . | . | $n_{\bullet c}$ | $n_{\bullet\bullet} = N$ |

Esta tabla es conocida como tabla de contingencia de rxc. La celda *ij* representada por $n_{ij}$ corresponde a la frecuencia observada en la categoría *i* de la variable 1 y la categoría *j* de la variable 2. El número total de observaciones en la categoría *i* de la variable 1 se denota por $n_{i\bullet}$ y son conocidos como totales marginales y lo mismo para $n_{\bullet j}$. Estos totales marginales están dados por:

$$n_{i\bullet} = n_{i1} + n_{i2} + \ldots\ldots + n_{ic} = \sum_{j=1}^{c} n_{ij}$$

$$n_{\bullet j} = n_{1j} + n_{2j} + \ldots\ldots + n_{rj} = \sum_{i=1}^{r} n_{ij}$$

$$n_{\bullet\bullet} = \sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij}$$

donde $n_{\bullet\bullet}$ representa el número total de observaciones ($N$).

El punto principal es si estas variables cualitativas que forman la tabla de contingencia son independientes o no.

Suponiendo que $p_{ij}$ representa la probabilidad de que una observación pertenezca a la celda *ij*, entonces la frecuencia $F_{ij}$ que se espera en dicha celda es:

$$F_{ij} = Np_{ij}$$

$$F_{ij} = E(n_{ij})$$

esto suponiendo que las frecuencias observadas siguen una distribución multinomial[12] .

Entonces por la ley de multiplicación de probabilidades para dos variables independientes entre si se tiene que:

$$\Rightarrow F_{ij} = Np_{i\bullet}p_{\bullet j}$$

Estas probabilidades pueden ser estimadas usando el método de máxima verosimilitud obteniendo,

$$\hat{p}_{i\bullet} = n_{i\bullet}/N \quad \text{y} \quad \hat{p}_{\bullet j} = n_{\bullet j}/N$$

$$\Rightarrow E_{ij} = N\hat{p}_{i\bullet}\hat{p}_{\bullet j} = Nn_{i\bullet}n_{\bullet j}/NN = n_{i\bullet}n_{\bullet j}/N \qquad (3.1)$$

[12] Canavos, G C., *Probabilidad y Estadística*, Ed. Mc. Graw Hill. Pag.186,187

Cuando dos variables son independientes las frecuencias estimadas usando (3.1) y las frecuencias observadas difieren muy poco; sin embargo, en el caso que las dos variables no sean independientes habrá grandes diferencias.

Para probar independencia se necesita ver si se rechaza o se acepta la hipótesis nula ($H_0$) y para esto se utiliza la prueba sugerida por Pearson, la prueba $X^2$ que está dada por:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (3.2)$$

Se puede ver que la magnitud de esta estadística depende de el valor de $(n_{ij} - E_{ij})$, en el cual se ve si existe diferencia entre las frecuencias que se observan y las correspondientes que se esperan, tal que la hipótesis nula de independencia entre variables se rechace.

La estadística $X^2$ se aproxima a una distribución chi-cuadrada, una forma de probar la hipótesis nula es comparando el valor $X^2$ con los valores tabulados de esta distribución, usando rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1) grados de libertad y un nivel de significancia de 0.05, o el que se desee.

Las dos modalidades que se mencionaron anteriormente son el análisis para **tablas de contingencia multidimensionales** y los modelos log-lineales. El primero es casi análogo al de una tabla de dos dimensiones, la diferencia con éste, radica en las hipótesis sugeridas; en el caso de mutua independencia en una tabla de dos dimensiones es como sigue:

$$H_0 : p_{ij} = p_{i\bullet}p_{\bullet j} \qquad \text{para } i=1, 2, \ldots, r;$$

$$j=1, 2, \ldots, c.$$

mientras que en las tablas de tres dimensiones por ejemplo la hipótesis de mutua independencia es:

$$H_0 : p_{ijk} = p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k} \qquad \text{para } i=1, 2, \ldots, r$$

$$j=1, 2, \ldots, c$$

$$k=1, 2, \quad , l$$

pero en este caso si la prueba es significativa no se puede asumir que la independencia es entre todas las variables. Puede ser independencia parcial o independencia condicional. Las siguientes hipótesis son para independencia parcial:

$$(1)\ H_0 : p_{ijk} = p_{i\bullet\bullet}p_{\bullet jk}$$

$$(2)\ H_0 : p_{ijk} = p_{\bullet j\bullet}p_{i\bullet k}$$

$$(3)\ H_0 : p_{ijk} = p_{\bullet\bullet k}p_{ij\bullet}$$

en (1) si la hipótesis se acepta esto implica que *i* es independiente de *j* y *k*, es decir,

$$p_{ij\bullet} = p_{i\bullet\bullet}p_{\bullet j\bullet} \quad \text{y} \quad p_{i\bullet k} = p_{i\bullet\bullet}p_{\bullet\bullet k}$$

$$E_{ijk}=N\hat{p}_{i\bullet\bullet}\hat{p}_{\bullet jk}$$

donde $$\hat{p}_{i\bullet\bullet}=\frac{n_{i\bullet\bullet}}{N} \quad \text{y} \quad \hat{p}_{\bullet jk}=\frac{n_{\bullet jk}}{N}$$

$$\Rightarrow E_{ijk}=\frac{n_{i\bullet\bullet}n_{\bullet jk}}{N}$$

usando como grados de libertad rcl - (r - 1) - (cl - 1) - 1 = rc - r - cl + 1 para el caso de independencia parcial.

Por último se tienen los **modelos log-lineales**, en éste se trata de encontrar un modelo lineal adecuado que muestre cuales son las variables o las combinaciones de ellas que están determinando los valores tomados por las observaciones; dichos modelos postulan que los valores esperados de las observaciones están dados por una combinación lineal de los parámetros, donde los parámetros representan los principales efectos de las variables. Se habla entonces de "interacción" como una alternativa al término asociación, es decir, interacción de primer orden entre pares de variables, interacción de segundo orden entre tres variables, etc.

Partiendo de la hipótesis de independencia para tablas de dos dimensiones, y tomando ahora los logaritmos naturales, de esta ecuación se tiene:

$$\log_e p_{ij} = \log_e p_{i\bullet} + \log_e p_{\bullet j} \tag{3.3}$$

si se considera que $F_{ij} = Np_{ij} \Rightarrow p_{ij} = \frac{F_{ij}}{N}$, entonces la ecuación (3.3) se puede escribir en términos de frecuencias de la siguiente forma:

$$\log_e \frac{F_{ij}}{N} = \log_e \frac{F_{i\bullet}}{N} + \log_e \frac{F_{\bullet j}}{N}$$

$$\Rightarrow \log_e F_{ij} - \log_e N = \log_e F_{i\bullet} - \log_e N + \log_e F_{\bullet j} - \log_e N$$

$$\Rightarrow \log_e F_{ij} = \log_e F_{i\bullet} + \log_e F_{\bullet j} - \log_e N$$

sumando sobre *i* y *j*, entonces queda:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{i\bullet} + \sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{\bullet j} - \sum_{i=1}^{r}\sum_{j=1}^{c}\log_e N$$

$$\Rightarrow \sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{ij} = c\sum_{i=1}^{r}\log_e F_{i\bullet} + r\sum_{j=1}^{c}\log_e F_{\bullet j} - rc\ \log_e N$$

$$\Rightarrow \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{ij}}{rc} = \frac{c\sum_{i=1}^{r}\log_e F_{i\bullet} + r\sum_{j=1}^{c}\log_e F_{\bullet j} - rc\log_e N}{rc}$$

$$\Rightarrow \log_e N = \frac{\sum_{i=1}^{r}\log_e F_{i\bullet}}{r} + \frac{\sum_{j=1}^{c}\log_e F_{j\bullet}}{c} - \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{ij}}{rc}$$

esto se puede expresar de la siguiente manera:

$$\log_e F_{ij} = u + u_{1(i)} + u_{2(j)} \tag{3.4}$$

donde

$$u = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{ij}}{rc}$$

$$u_{1(i)} = \frac{\sum_{j=1}^{r}\log_e F_{ij}}{c} - \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{ij}}{rc} \tag{3.5}$$

$$u_{2(j)} = \frac{\sum_{j=1}^{c}\log_e F_{ij}}{r} - \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\log_e F_{ij}}{rc} \tag{3.6}$$

Se puede ver que (3.4) especifica un modelo lineal para el logaritmo de las frecuencias, es decir, un modelo log-lineal. Donde el parámetro $u$ representa el efecto general, la $u_{1(i)}$ representa el principal efecto en la categoría $i$ de la variable 1 y la $u_{2(j)}$ representa el principal efecto en la categoría $j$ de la variable 2.

Examinando las ecuaciones (3.5) y (3.6) se puede ver que los parámetros $u_{1(i)}$ y $u_{2(j)}$ son medidos por la variación entre una columna o fila con el efecto general, es decir:

$$u_{1(\bullet)} = 0 \quad \text{y} \quad u_{2(\bullet)} = 0$$

Ahora, lo que interesa es extender el modelo especificado en (3.4) a la situación en la cual las variables son independientes, para esto se introduce el término que representa la interacción entre dos variables, dado por:

$$\log F_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{3.7}$$

donde $u_{12(ij)}$ representa la interacción entre las categorías $i$ y $j$ de las variables 1 y 2 respectivamente, donde:

$$u_{12(i\bullet)} = 0 \quad \text{y} \quad u_{12(\bullet j)} = 0$$

La estimación del efecto de interacción se utilizará para identificar las categorías responsables de cualquier desviación de independencia.

Probar independencia es equivalente a probar si los términos de interacción en (3.6) son cero, por ejemplo la hipótesis de no-interacción de segundo orden es como sigue:

$$H_0 : u_{123} = 0 \qquad \forall \text{i, j, k}$$

El proceso es similar a los anteriores, es decir, obteniendo estimadores de las frecuencias teóricas que se esperan $(E_{ijk})$, asumiendo que el modelo es correcto y comparando esto con los valores observados promedios de $X^2$. Si se toma,

$$z_{ij} = \log E_{ij}$$

$$\Rightarrow z_{i\bullet} = \frac{1}{c}\sum_{j=1}^{c} \log E_{ij}$$

por lo tanto se toma como $\hat{u} = \bar{z}_{\bullet\bullet}$ y si se extiende a tres variables se tendrá que $\hat{u} = \bar{z}_{\bullet\bullet\bullet}$, entonces

$$\hat{u}_{1(i)} = \bar{z}_{i\bullet\bullet} - \bar{z}_{\bullet\bullet\bullet}$$

$$\hat{u}_{2(j)} = \bar{z}_{\bullet j\bullet} - \bar{z}_{\bullet\bullet\bullet}$$

$$\hat{u}_{3(k)} = \bar{z}_{\bullet\bullet k} - \bar{z}_{\bullet\bullet\bullet}$$

$$\hat{u}_{23(jk)} = \bar{z}_{\bullet jk} - \bar{z}_{\bullet j\bullet} - \bar{z}_{\bullet\bullet k} - \bar{z}_{\bullet\bullet\bullet\bullet}$$

de esta misma forma se puede extender a más de tres variables.

Al inicio de la investigación no se contaba con los datos, únicamente se tenía el conocimiento de la estructura de los mismos, por lo que, este último análisis se consideraba como el adecuado para este estudio en específico. Sin embargo, no fue posible aplicarlo por dos razones principalmente, la primera es que no se podía determinar cuales genes provenían del padre y cuales de la madre, lo cual era esencial para construir esta tabla. El segundo problema que ya se mencionó, fue el tamaño de la muestra, ya que si se considera que cada individuo aporta dos combinaciones genéticas de un conjunto de alrededor de 3,000 millones de posibles combinaciones, la presencia de ceros en la tabla sería demasiada, aún suponiendo que se eliminarían muchas columnas.

### 3.3.4.2 Regresión Logística

Otro método fuera del análisis de tablas de contingencia que se estudió y se aplicó fue el de regresión logística. Donde se intenta encontrar un modelo que describa la relación que existe entre una variable dependiente o respuesta y un conjunto de variables independientes, como en otros modelos de regresión. Sin embargo, lo que distingue al modelo de regresión logística con el lineal, es que la variable respuesta en el primero es binaria o dicotómica mientras que en el segundo el continua.

En el caso particular de este estudio se tienen tres variables respuesta y una variable independiente por cada uno de los genes considerados, lo que implica un tratamiento especial, sin embargo, a manera de introducción de este método se considerará el caso en que se tiene una variable respuesta y una independiente, es decir, **regresión logística simple**.

Para el caso de Regresión lineal, si se grafican las coordenadas de las observaciones, generalmente se dibuja una cierta tendencia, pero tomando el caso de regresión logística todos los puntos quedan concentrados en las rectas $y$=0 y $y$=1 o en los mismos puntos, lo que impide describir una relación funcional entre variables. Para suavizar esta variabilidad sin romper con la estructura de la relación, se forman grupos en la variable independiente, y la variable respuesta será el valor medio de los datos observados en cada grupo, media condicional $E(Y/x)$. En regresión lineal ésta se puede expresa como una ecuación lineal en $x$ de la siguiente manera:

$$E(Y/x) = \beta_0 + \beta_1 x$$

esta expresión implica que $E(Y/x)$ puede tomar cualquier valor para x entre los rangos $-\infty$ y $+\infty$. Para datos dicotómicos como es el caso del problema de regresión logística la media condicional es mayor o igual a 0 y menor o igual a 1. El modelo que se usa para $E(Y/x)$ en el caso en que *y* es dicotómica, es el de la distribución logística, debido a que es una función muy flexible y fácil de usar, y otra razón es que otorga una interpretación biológicamente significativa, Cox (1970) discutió sobre esto.

Para simplificar la notación, se usará $\pi(x) = E(Y/x)$ para representar la media condicional cuando la distribución logística es utilizada. La forma específica del modelo de regresión logística es como sigue:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

La transformación logit de $\pi(x)$ está definida en términos de $\pi(x)$ como sigue,

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

la importancia de esta transformación es que $g(x)$ tiene muchas de las propiedades deseables de un modelo de regresión lineal El modelo logit, $g(x)$ es lineal en sus parámetros, puede ser continuo y puede tomar rangos de $-\infty$ a $\infty$, dependiendo del rango de x. Otra diferencia con respecto a la regresión lineal es el método utilizado para estimar parámetros desconocidos, uno es el de mínimos cuadrados, sin embargo, este método de estimación cuando la variable dependiente es dicotómica no mantiene las mismas propiedades, por lo que, se utiliza el de máxima verosimilitud.

Es fácil extender el método para el caso en el que hay más de una variable independiente, denominado como **regresión logística múltiple**. El logit de este modelo esta dado por la ecuación,

$$g(x) = \beta_0 + \beta_1 x_1 + ..... + \beta_m x_m$$

donde el vector $x = (x_1, x_2, ....., x_m)$ es la colección de m variables independientes y

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Existe otro caso denominado **regresión logística politómica o multinomial**, el cual es el caso particular de este estudio, donde se cuenta con tres variables respuesta ($y = 1$, $y = 2$ y $y = 3$), donde la respuesta 1 no implica que es menor que la respuesta 2 ó 3. Esta es una propiedad que distingue el modelo logit multinomial del modelo logit, ya que este último cuenta con dos respuestas 0 ó 1, donde 0 implica ausencia y 1 presencia.

En el modelo logit multinomial se desea estimar un conjunto de coeficientes $\beta^{(1)}$, $\beta^{(2)}$ y $\beta^{(3)}$ correspondientes a cada una de las respuestas. Para poder identificar plenamente el modelo es necesario tomar una de las tres como base, para este estudio se considerará la categoría CONTROLES, es decir, $\beta^{(1)} = 0$, entonces los coeficientes $\beta^{(2)}$ y $\beta^{(3)}$ serán medidos con respecto al grupo $y = 1$. De lo contrario se tendrían más de una solución para $\beta^{(1)}$, $\beta^{(2)}$ y $\beta^{(2)}$ con las mismas probabilidades para $y = 1$, $y = 2$ y $y = 3$.

Tomando $\beta^{(1)} = 0$, entonces

$$P(y=1) = \frac{1}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}}$$

$$P(y=2) = \frac{e^{X\beta^{(2)}}}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}}$$

$$P(y=3) = \frac{e^{X\beta^{(3)}}}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}}$$

La probabilidad relativa con respecto a la categoría base es denominada como riesgo relativo y está dado por:

$$\frac{P(y=i)}{P(y=1)} = e^{X\beta^{(i)}} \qquad \text{Para i = 2 ó 3}$$

El riesgo relativo para el cambio de una unidad en $x_i$ es entonces:

$$\frac{e^{\beta_1^{(2)}x_1 + \quad + \beta_i^{(2)}(x_i+1) + \quad + \beta_m^{(2)}x_k}}{e^{\beta_1^{(2)}x_1 + \quad + \beta_i^{(2)}x_i + \quad + \beta_m^{(2)}x_k}} = e^{\beta_i^{(2)}}$$

# CAPÍTULO 4

## 4.1 Introducción

En el capítulo anterior se dio una descripción de dos métodos de análisis de datos que cumplen con el objetivo buscado en el presente estudio. El primero es el análisis de tablas de contingencia bidimensionales, con éste se busca poder detectar cuáles genes son los que más se desvían de su valor esperado, es decir, que su presencia o ausencia en cierto grupo (DM, DMCRD o DMSRD) es notable con respecto a la frecuencia del grupo control (CONTROLES). El segundo método mencionado es el de regresión logística, con el que se pretende obtener un modelo mediante el cual se pueda determinar la probabilidad que tiene un individuo de desarrollar la enfermedad, suponiendo que la muestra con la que se cuenta puede representar a la población general. En este capítulo se presentarán los resultados obtenidos de la aplicación de estos métodos.

## 4.2 Análisis de Tablas de Contingencia utilizando la Prueba ji-cuadrada

Inicialmente se formaron cinco tablas de contingencia para cada una de las variables A, B, C, DR y DQ, tomando para las columnas las categorías de éstas y para las filas las posibles combinaciones de las tres categorías de la variable

enfermedad, las cuales se presentan como criterios a continuación, con el fin de mencionarlas de esta forma más adelante,

1er. criterio: controles y diabéticos (DMCRD y DMSRD).
2do. criterio: controles, diabéticos con retinopatía y diabéticos sin retinopatía.
3er. criterio: diabéticos con retinopatía y diabéticos sin retinopatía.
4to. criterio: controles y diabéticos sin retinopatía.
5to criterio: controles y diabéticos con retinopatía.

Esto con la finalidad de que al obtener el valor de la estadística de prueba para cada una de estas tablas, si éste pertenece a la región crítica de la prueba , es decir, si el valor es mayor o igual que el valor crítico entonces se rechazará la hipótesis nula de independencia, con esto se tendrá una idea muy general de cuales variables son las más influyentes en el desarrollo de estas enfermedades, aunque el interés de este trabajo es específicamente la retinopatía diabética.

Otra consideración que es importante mencionar, es que derivado de que cada una de las personas aporta dos genes de cada variable A, B, C, DR y DQ, puede suceder que el alelo se repita o que sean distintos, por lo que, se realizó un análisis para los datos completos (total), es decir, si algún individuo tenía Ax y Ay donde x=y entonces se incluían las dos, y un análisis eliminando x o y indistintamente (parcial), ya que para el caso en el que $x \neq y$ no se podía eliminar uno de los dos, debido a que no existía un criterio válido de eliminación. Si se supiera cual gen provenía del padre o cual de la madre, tal vez se hubiera tomado otra decisión.

En la tabla T.4.1 se presentan los resultados mencionados anteriormente con su respectivo valor crítico. Como se puede ver para A, B, C y DQ no hay ninguna evidencia de asociación, en todos los casos se aceptaría la hipótesis nula de independencia. Sin embargo, para la variable DR en la mitad de los casos se rechaza la hipótesis nula.

**Tabla T.4.1.** Valores obtenidos de aplicar la estadística de prueba a las tablas generales.

| | | A | B | C | DR | DQ |
|---|---|---|---|---|---|---|
| 1er.criterio | Total | 13.3895 | 17.2269 | 14.0004 | 18.8877 | 5.9073 |
| | Parcial | 11.7613 | 14.8550 | 10.3462 | 16.5702 | 5.3545 |
| 2do.criterio | Total | 15.2075 | 29.7496 | | 34.3977 | 7.3245 |
| | Parcial | 13.6329 | 23.2521 | | 31.5096 | 6.3591 |
| 3er.criterio | Total | 1.3582 | 13.4390 | | 10.1351 | 1.5248 |
| | Parcial | 2.0400 | 7.9655 | | 9.5120 | 1.0252 |
| 4to.criterio | Total | 3.4550 | 12.6984 | | 17.6868 | 2.0802 |
| | Parcial | 3.4945 | 7.8163 | | 17.6249 | 2.6041 |
| 5to.criterio | Total | 12.7444 | 16.3894 | 12.2263 | 15.2125 | 5.7728 |
| | Parcial | 11.0250 | 15.0258 | 9.4788 | 12.5567 | 4.6109 |
| 1,3,4 y 5 criterios | Valor | 14.07 | 27.59 | 15.51 | 16.92 | 7.82 |
| 2do.criterio | crítico | 23.69 | 49.81 | | 28.87 | 12.59 |

**Nota.** El valor crítico de los criterios se separaron debido a que el 2do criterio tiene 3 categorías en las columnas y por lo tanto, los grados de libertad aumentan.

También se puede ver que la diferencia en las dos consideraciones (total y parcial), es que la primera casi en todos los casos es simplemente mayor que la segunda, por lo que puede ocurrir que, si tomando los datos completos la hipótesis nula se rechazaba, con la segunda consideración no.

Para el caso de la variable C, no se formaron todas las tablas debido a que sólo se contaba con los datos de 12 de los pacientes diabéticos, de los cuales sólo dos no presentan retinopatía diabética, por esta razón se realizó el análisis del 1er. y el 5to. criterio únicamente.

Es un buen adelanto, saber de manera general cuáles son las variables importantes en el desarrollo de las enfermedades en cuestión. Sin embargo, el propósito de este análisis es determinar específicamente qué genes son los más significativos, por ejemplo, de la tabla T.4.1 se puede decir que los genes del tipo DR podrían estar asociados con la retinopatía diabética, sin embargo, no se sabe qué genes son los que están ocasionando esto.

Para resolver lo anterior, se formaron tablas de contingencia de 2x2 y para el 2do. criterio de 3x2, donde las filas mantienen igual sus categorías y las columnas se dividen en dos únicamente, la primera contiene la frecuencia de un sólo gen y la segunda es la frecuencia total excluyendo dicho gen, con esto los totales marginales no se modifican.

Considerando primero los resultados para la variable DR, únicamente se tomaron los resultados en donde el valor crítico para $\alpha = 0.05$ de las dos consideraciones llevó a rechazar la hipótesis nula. Estos valores se pueden observar en la tabla T.4.2, es decir, el gen DR3 y DR6 en el 1er. criterio, el DR2 y DR9 en el 2do. criterio, en el 3er. criterio ninguno resultó importante, en el 4to. criterio el DR6 y DR9, y por último en el 5to. criterio el DR2. Analizando cada uno de estos, se pueden eliminar algunos, ya que si se toma primero el resultado obtenido en el 1er. criterio, el gen DR3 no se repite en ningún otro criterio, por lo que, la probable asociación de este gen es con la diabetes mellitus. Sin embargo, el resultado para el gen DR6 está estrechamente relacionado con el obtenido en el 4to. criterio con respecto al mismo gen, por lo tanto, se tomó únicamente el segundo. Si se procede de la misma forma con todos los que están sombreados, quedarán como significativos los siguientes,

| | |
|---|---|
| 4to. criterio (CONTROLES y DMSRD) | DR6, DR9 |
| 5to. criterio (CONTROLES y DMCRD) | DR2 |

Con ninguna otra variable A, B, C y DQ, los resultados de la tabla T.4.1 lleva a pensar que existe alguna asociación, sin embargo, se realizó la misma separación de las tablas generales para éstas.

Considerando ahora la tabla T.4.3 donde se muestran los valores resultantes de la subdivisión de la variable B. Como se puede ver, si hay un caso en el que se rechaza la hipótesis nula, para B40 existe una posible asociación en el 2do. criterio, donde el resultado de la tabla total rechaza la hipótesis nula y el de la tabla parcial la acepta, tomando el mismo criterio de únicamente considerar los valores en donde ambas consideraciones rechazan la hipótesis nula, éste quedaría descartado, sin embargo, esta posible asociación está explicada en el 4to. criterio, por lo que, si hay otro resultado que se debe considerar y es el siguiente,

4to. criterio (CONTROLES y DMSRD) B40

Para el caso de las variables DQ y A únicamente se encontró un caso en cada una en donde la primera consideración rechazó la hipótesis nula, sin embargo, en la segunda consideración no, por lo que no se consideró ningún resultado como importante.

Por último en la variable C, el gen C3 resultó significativo en en 1er. criterio, tanto en la consideración total como en la parcial, sin embargo, realizando el mismo análisis para el 5to. criterio en la consideración (total) se rechaza la hipótesis nula y en la otra (parcial) se acepta. Debido a lo anterior otro resultado debería incluirse a la lista, el gen C3 en el 1er. criterio, sin embargo, como ya se ha dicho el tema principal de este trabajo es la retinopatía, y como no existe ninguna evidencia de que el gen C3 esté involucrado con ésta, no se considerará como resultado.

Para todo lo anterior se tomó en cuenta como máximo una $\alpha = 0.05$[13], ya que por ejemplo con $\alpha = 0.10$ se hubieran obtenido más resultados, pero debido a las características de la muestra que por un lado es pequeña y por otro lado, no es aleatoria, se fue muy estricto para tomar en cuenta sólo los que tuvieran la una baja probabilidad de error.

Las tablas que se presentan a continuación son las correspondientes a las frecuencias de los genes DR2, DR6, DR9 y B40 con su respectivo valor esperado encerrado en un paréntesis.

| | CDR2 | SDR2 |
|---|---|---|
| CONTROLES | 3(6.89) | 86(82.11) |
| DMCRD | 8(4.11) | 45(48.89) |

| | CDR6 | SDR6 |
|---|---|---|
| CONTROLES | 6(8.48) | 83(80.52) |
| DMSRD | 4(1.72) | 12(14.48) |

| | CDR9 | SDR9 |
|---|---|---|
| CONTROLES | 0(0.85) | 89(88.15) |
| DMSRD | 1(0.15) | 15(15.85) |

| | CB40 | SB40 |
|---|---|---|
| CONTROLES | 8(9.52) | 71(69.48) |
| DMSRD | 2(0.48) | 2(3.52) |

donde CDR2 quiere decir con el gen DR2 y SDR2 sin el gen respectivo.

Entonces si la frecuencia observada es mayor a la esperada, implica que la presencia de este gen podría estar asociada con la presencia o ausencia de la enfermedad, según del caso que se trate, e inversamente, si la frecuencia observada es menor que la esperada, existe una ausencia notable del gen en cuestión.

[13] Probabilidad de cometer el error de tipo I = P(rechazar $H_0 | H_0$ es cierta) = $\alpha$

**Tabla T.4.2.** Valores obtenidos de aplicar la estadística de prueba a las subtablas de la variable DR.

| | | DR1 | DR2 | DR3 | DR4 | DR5 | DR6 | DR7 | DR8 | DR9 | DR10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1er criterio | Total | 0.2899 | 4.0578 | 4.6074 | 1.9835 | 1.2924 | 4.3669 | 1.1942 | 1.7869 | 1.2981 | 0.2709 |
| | Parcial | 0.5688 | 2.9090 | 4.0405 | 2.3918 | 0.8818 | 4.2352 | 1.2697 | 1.0817 | 1.2715 | 0.0372 |
| 2do criterio | Total | 1.4619 | 8.3807 | 4.8392 | 2.7049 | 1.4088 | 5.5615 | 1.5582 | 1.8202 | 8.9315 | 0.9114 |
| | Parcial | 1.2798 | 7.1584 | 4.3342 | 2.5751 | 0.9390 | 4.7780 | 1.6966 | 1.1879 | 7.6807 | 0.5855 |
| 3er criterio | Total | 1.0406 | 2.7318 | 0.6218 | 0.8059 | 0.1812 | 0.8394 | 0.6218 | 0.0268 | 3.3612 | 0.8313 |
| | Parcial | 0.6044 | 2.8115 | 0.7352 | 0.2164 | 0.0823 | 0.3896 | 0.7352 | 0.0879 | 2.8594 | 0.6038 |
| 4to criterio | Total | 1.3778 | 0.5552 | 2.2089 | 2.2487 | 0.1298 | 5.2469 | 1.1440 | 0.9996 | 5.6160 | 0.0922 |
| | Parcial | 1.2879 | 0.6442 | 2.3283 | 1.6136 | 0.1360 | 4.0876 | 1.3327 | 0.8879 | 4.8648 | 0.1783 |
| 5to criterio | Total | 0.0079 | 6.3886 | 2.9444 | 0.9351 | 1.3801 | 2.6078 | 0.5504 | 1.4159 | 0 | 0.6649 |
| | Parcial | 0.1307 | 5.1310 | 2 3370 | 1.4938 | 0.8974 | 2.7878 | 0.5195 | 0.7029 | 0 | 0.2509 |
| 1,3,4 Y 5 criterios | Valor | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 |
| 2do.criterio | crítico | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 |

**Tabla T.4.3.** Valores obtenidos de aplicar la estadística de prueba a las subtablas de la variable B.

| | | B5 | B7 | B8 | B12 | B14 | B15 | B16 | B17 | B18 | B21 | B22 | B27 | B35 | B37 | B40 | B41 | B48 | B70 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1er criterio | Total | 0.0258 | 0.3332 | 0.0341 | 1.1966 | 2.9647 | 0.0534 | 1.53 | 0.4573 | 3.4169 | 0.7412 | 2.52 | 0.4032 | 0.2250 | 0.8129 | 1.0954 | 0.8129 | 0.4032 | 1.2293 |
| | Parcial | 0.0398 | 0.2795 | 0.0222 | 0.2677 | 1.7328 | 0.0826 | 0.58 | 0.4132 | 3.5988 | 0.6444 | 2.4155 | 0.4215 | 0.0131 | 0.8506 | 1.3809 | 0.8506 | 0.4215 | 1.2877 |
| 2do.criterio | Total | 0.2066 | 0.9164 | 0.2732 | 2.5368 | 2.9647 | 1.5483 | 2.7199 | 0.8129 | 3.4169 | 1.6941 | 3.2256 | 0.4032 | 1.8 | 0.8129 | 6.5852 | 0.8129 | 0.4032 | 1.2293 |
| | Parcial | 0.1611 | 0.6715 | 0.1825 | 0.8348 | 1.7328 | 2.5442 | 0.7528 | 0.6515 | 3.5988 | 1.2864 | 2.8879 | 0.4215 | 1.2682 | 0.8506 | 5.5941 | 0.8506 | 0.4215 | 1.2877 |
| 3er criterio | Total | 0.2057 | 0.4235 | 0.2057 | 0.9 | | 1.6941 | 0.8867 | 0.2057 | | 0.6789 | 0.2057 | | 1.7379 | | 4.2916 | | | |
| | Parcial | 0.1422 | 0.2937 | 0.1422 | 0.4552 | | 2.8678 | 0.1422 | 0.1422 | | 0.4552 | 0.1422 | | 1.2255 | | 3.0977 | | | |
| 4to criterio | Total | 0.2065 | 0.2065 | 0.1362 | 0.3516 | 0.5034 | 0.8321 | 2.5397 | 0.0674 | 0.5818 | 0.2783 | | 0.0674 | 1.7920 | 0.1362 | 6.4446 | 0.1362 | 0.0674 | 0.2065 |
| | Parcial | 0.1576 | 0.1576 | 0.1038 | 0.2694 | 0.2128 | 1.4935 | 0.5044 | 0.0513 | 0.4483 | 0.2128 | | 0.0513 | 1.0036 | 0.1038 | 5.7123 | 0.1038 | 0.0513 | 0.1576 |
| 5to criterio | Total | 0 | 0.6261 | 0.1140 | 1.9620 | 2.4779 | 0.4551 | 0.6349 | 0.6780 | 2.8571 | 1.2642 | 3.0252 | 0.3361 | 0 | 0.6780 | 0.0254 | 0.6780 | 0.3361 | 1 0256 |
| | Parcial | 0.0073 | 0.4614 | 0.0660 | 0.4988 | 1.5248 | 0.6018 | 0.3726 | 0.5559 | 3.1716 | 0.9762 | 2.7496 | 0.3705 | 0.1910 | 0 7480 | 0.2887 | 0.7480 | 0.3705 | 1.1327 |
| 1,3,4 y 5 criterios | Valor | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 | 3 84 | 3.84 | 3.84 | 3.84 | 3.84 | 3.84 |
| 2do.criterio | crítico | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 |

## 4.3 Regresión Logística

Para realizar este análisis se formó inicialmente una tabla global, de la siguiente manera:

| | A1..... | B1..... | C1..... | DR1..... | DQ1..... | Posibles respuestas |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | CONTROLES = 1 |
| 3 | | | | | | |
| | | | 0's y 1's | | | DMCRD = 2 |
| | | | | | | |
| | | | | | | DMSRD = 3 |
| 85 | | | | | | |

donde la primera columna contienen a cada una de las personas que participaron en el estudio numeradas de forma consecutiva. En la primera fila se encuentran todos y cada uno de los genes considerados, éstos representan a las variables independientes donde las celdas intermedias contienen valores de 0 y 1, donde 1 significa la presencia del gen en la persona y 0 la ausencia del mismo. Por otro lado, la última columna contiene las tres respuestas: CONTROLES, DMCRD y DMSRD, a las cuales se les asignó los números 1, 2 y 3, respectivamente.

Para los casos donde la persona presentaba dos genes iguales, se eliminó uno de éstos, equivalente a la consideración parcial del análisis de tablas de contingencia.

Se utilizó un paquete estadístico (STATA versión 3.0) para obtener los coeficientes $\beta^{(2)}$ y $\beta^{(2)}$, ya que como se mencionó en el capítulo anterior el grupo control fue utilizado como grupo base y por lo tanto $\beta^{(1)} = 0$. Esto con la finalidad de obtener las siguientes probabilidades predictivas:

$$P(y=1) = \frac{1}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}}$$

$$P(y=2) = \frac{e^{X\beta^{(2)}}}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}}$$

$$P(y=3) = \frac{e^{X\beta^{(3)}}}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}}$$

asumiendo que $X$ y $\beta_k^{(i)}$ son vectores iguales a $(x_1, x_2, \ldots\ldots, x_k)$ y $\left(\beta_1^{(i)}, \beta_2^{(i)}, \ldots\ldots, \beta_k^{(i)}\right)$, respectivamente, donde i=2, 3.

No se pudo correr el paquete estadístico con la matriz original, debido al tamaño de ésta, con esto se hubiera obtenido el modelo general que incluía a todos y cada uno de los genes considerados con sus respectivos coeficientes, de cualquier forma a partir de este modelo general se iba a obtener el modelo ajustado utilizando algún método de eliminación de variables independientes. Derivado de lo anterior se utilizaron dos métodos distintos de ajuste para asegurar que el resultado fuera el mismo, en los dos se obtuvieron las mismas variables y los mismos coeficientes para estas variables.

El primero de éstos fue considerando todas las variables A, B y C dentro de un grupo y obteniendo los coeficientes de éstas, a partir de esta corrida se fueron eliminando variables utilizando como criterio los casos donde la probabilidad de que P>I t I era grande en las repuestas 2 y 3, esta probabilidad es equivalente al

error de tipo I, es decir, rechazar la hipótesis nula dado que es cierta, donde la estadística de prueba t tiene una distribución t de Student con n - 2 grados de libertad[14]. Posteriormente se consideraron las variables DR y DQ en otro grupo y se realizó el mismo proceso anterior.

Habiendo disminuido el número de variables independientes se juntaron los dos grupos y se corrió el programa siguiendo el mismo proceso, hasta quedar únicamente con las variables que se consideraron dentro del modelo.

El otro método de eliminación fue tomando todas las variables A, B, C, DR y DQ de manera independiente, eliminando una por una según el valor de la suma de las probabilidades de las repuestas 2 y 3, habiendo disminuido las variables se procedió a juntarlas de nuevo y continuar el proceso hasta obtener el mismo resultado que con el método de eliminación anterior.

En el apéndice 1 se muestra el resultado de tres corridas, una anterior a la definitiva (1), la definitiva (2) y una posterior a ésta (3), en la corrida (3) se puede ver la razón de porque no se continuó depurando[15].

También se realizó la tabla de contingencia para estas tres corridas obteniendo los siguientes valores:

| | | | |
|---|---|---|---|
| (1) | $X^2 = 18.008$ | con 10 grados de libertad | P = 0.05483 |
| (2) | $X^2 = 15.098$ | con 8 grados de libertad | P = 0.05726 |
| (3) | $X^2 = 14.810$ | con 6 grados de libertad | P = 0.02179 |

[14] Canavos, G. C., *Probabilidad y Estadística*, Ed. McGraw-Hill, Pag. 465-470.
[15] El valor de las probabilidades aumenta significativamente con respecto a la corrida anterior.

el valor de la P de la tabla de contingencia anterior a (1) fue de 0.10779 y de las anteriores a ésta es aún mayor, lo que lleva a pensar que escoger alguna de estas tres opciones sería una buena decisión, ya que los valores de P de (1), (2) y (3) son razonables para decidir rechazar la hipótesis nula de independencia.

La corrida seleccionada como definitiva fue la (2), esto fue porque a pesar de tener un valor de P menor en la corrida (1) para las tablas de contingencia, las probabilidades de cometer el error de tipo I para la variable A2 eran grandes y si se excluye dicha variable, estas probabilidades para las variables restantes disminuye aún más, lo contrario de lo que ocurre al eliminar la variable DQ1, como se puede ver en el apéndice 1 las probabilidades aumentan drásticamente.

Utilizando los coeficientes de la corrida (2) del apéndice 1 se tienen las siguientes probabilidades predictivas:

$$P(y=1)=\frac{1}{1+e^{-0.54-2.36A9+2.27DR1+2.36DR6+1.37DR8-2.09DQ1}+e^{-2.56-2.47A9+3.16DR1+3.29DR6+2.19DR8-1.71DQ1}}$$

$$P(y=2)=\frac{e^{-0.54-2.36A9+2.27DR1+2.36DR6+1.37DR8-2.09DQ1}}{1+e^{-0.54-2.36A9+2.27DR1+2.36DR6+1.37DR8-2.09DQ1}+e^{-2.56-2.47A9+3.16DR1+3.29DR6+2.19DR8-1.71DQ1}}$$

$$P(y=3)=\frac{e^{-2.56-2.47A9+3.16DR1+3.29DR6+2.19DR8-1.71DQ1}}{1+e^{-0.54-2.36A9+2.27DR1+2.36DR6+1.37DR8-2.09DQ1}+e^{-2.56-2.47A9+3.16DR1+3.29DR6+2.19DR8-1.71DQ1}}$$

Los modelos anteriores lo que muestran es el comportamiento general de las respuestas 1, 2 y 3, obtenido del comportamiento particular de la muestra observada. Para el mejor entendimiento de esto, se consideró un ejemplo y se aplicó a dichos modelos, éste fue suponiendo una persona que no contiene ninguno de los genes considerados en el modelo, es decir, si (A9, DR1, DR6, DR8, DQ1) = (0, 0, 0, 0, 0) entonces sustituyendo estos valores en los modelos se tiene las siguientes probabilidades predictivas,

$$P(y=1)=\frac{1}{1+e^{-0\,54}+e^{-2\,56}}=0.602$$

$$P(y=2)=\frac{e^{-0\,54}}{1+e^{-0\,54}+e^{-2\,56}}=0.3509$$

$$P(y=3)=\frac{e^{-2\,56}}{1+e^{-0\,54}+e^{-2\,56}}=0.0470$$

Estas serían las probabilidades que tiene una persona que está sana de desarrollar alguna de estas enfermedades, es decir, una persona que está sana y que carece de los genes A9, DR2, DR6, DR8 y DQ1, con una probabilidad de 0.6021 permanecerá sana, la probabilidad de desarrollar diabetes mellitus es mucho menor pero sigue siendo una probabilidad relativamente alta y por último es muy poco probable que habiendo desarrollado diabetes desarrolle después retinopatía.

Debido a que el modelo involucra únicamente cinco variables, las combinaciones de éstas son todos los posibles resultados de los modelos, por lo anterior se realizó la Tabla T.4.4, donde las primeras cinco columnas contienen las combinaciones mencionadas, los recuadros marcados con una cruz implican la

presencia de los genes, y las últimas tres columnas como se puede observar son las probabilidades de que *y* tome los valores de 1, 2 y 3 *obtenidas para cada uno* de los casos. Esta tabla sería la conclusión de la aplicación del método de regresión logística a los datos, ya que si se obtiene el código genético de una persona cualquiera se puede obtener la probabilidad de mantenerse o pertenecer a cualquiera de los tres grupos, esto suponiendo que la muestra puede representar a la población en general.

Cabe mencionar que se realizaron dos corridas adicionales, la primera con un propuesta que según los médicos involucrados en el presente estudio, son los genes que durante el tiempo y las investigaciones realizadas se han detectado como los determinantes en esta enfermedad, estos genes son los siguientes: A1, B8, B18, C7, DR3, DR4, DQ2 y DQ3, utilizando el mismo criterio de eliminación que en las anteriores, es decir, utilizando los valores P>I t I los cuales en casi todos los casos eran 1[16], por lo que, se empezó eliminando estos casos uno por uno, hasta que no quedó ninguno ya que en todos las probabilidades resultaron altas.

La segunda corrida mencionada fue utilizando los genes donde la estadística de prueba para las tablas de contingencia analizadas fue cercano al valor crítico correspondiente, estos genes fueron los siguientes A1, A9, A28, B18, B22, B40, C3, C7, DR2, DR3, DR4, DR6, DR8, DR9, DQ1, DQ2, DQ3 y DQ4, utilizando en mismo proceso de eliminación que para los casos anteriores, quedaron únicamente los genes A9, DR6.

[16] Esto implica que con una probabilidad de 1, se estaría rechazando la hipótesis nula de independencia siendo que es cierta.

**Tabla T.4.4** Posibles resultados de los modelos P(y=1), P(y=2) y P(y=3)

| A9 | DR1 | DR6 | DR8 | DQ1 | P(Y=1) | P(Y=2) | P(Y=3) |
|---|---|---|---|---|---|---|---|
| X | | | | X | 0.9921 | 0.0068 | 0.0012 |
| X | | | X | X | 0.9636 | 0.0258 | 0.0103 |
| X | | | | | 0.9419 | 0.0518 | 0.0062 |
| | | | | X | 0.9206 | 0.0664 | 0.0130 |
| X | X | | | X | 0.9141 | 0.0602 | 0.0257 |
| X | | X | | X | 0.9057 | 0.0653 | 0.0290 |
| X | | | X | | 0.7840 | 0.1698 | 0.0463 |
| | | | X | X | 0.7093 | 0.2012 | 0.0895 |
| X | X | | X | X | 0.6619 | 0.1716 | 0.1665 |
| X | | X | X | X | 0.6369 | 0.1807 | 0.1825 |
| | | | | | 0.6021 | 0.3509 | 0.0470 |
| X | X | | | | 0.5923 | 0.3155 | 0.0922 |
| X | | X | | | 0.5682 | 0.3311 | 0.1007 |
| | X | | | X | 0.4925 | 0.3436 | 0.1639 |
| | | X | | X | 0.4668 | 0.3563 | 0.1769 |
| X | X | X | | X | 0.4076 | 0.2844 | 0.3080 |
| | | | X | | 0.2506 | 0.5746 | 0.1748 |
| X | X | | X | | 0.2229 | 0.4671 | 0.3100 |
| X | | X | X | | 0.2050 | 0.4702 | 0.3248 |
| | X | | X | X | 0.1488 | 0.4086 | 0.4426 |
| | | X | X | X | 0.1353 | 0.4064 | 0.4583 |
| | X | | | | 0.1179 | 0.6651 | 0.2170 |
| | | X | | | 0.1079 | 0.6659 | 0.2262 |
| X | X | X | X | X | 0.0952 | 0.2615 | 0.6432 |
| X | X | X | | | 0.0924 | 0.5213 | 0.3862 |
| | X | X | | X | 0.0577 | 0.4265 | 0.5158 |
| | X | | X | | 0.0255 | 0.5600 | 0.4148 |
| | | X | X | | 0.0227 | 0.5518 | 0.4255 |
| X | X | X | X | | 0.0165 | 0.3667 | 0.6168 |
| | X | X | X | X | 0.0091 | 0.2646 | 0.7264 |
| | X | X | | | 0.0091 | 0.5424 | 0.4485 |
| | X | X | X | | 0.0015 | 0.3470 | 0.6515 |

# CONCLUSIONES

Como se menciona en el capitulo dos los pasos generales de una investigación son la observación empírica, la formulación de hipótesis, la identificación y control de las variables, la comprobación de dicha hipótesis y finalmente la postulación de leyes, modelos y teorías. Se dicen fácil pero llevar a término un proyecto de investigación obteniendo de éste resultados confiables, no es nada sencillo.

En el presente trabajo no hubo una observación empírica como tal, pero derivado de que la retinopatía diabética es una consecuencia de la diabetes mellitus, y esta última se ha comprobado que depende en parte de factores genéticos, entonces hay una razón para creer que la primera también, por lo tanto, la hipótesis del estudio es que probablemente existe alguna asociación entre el desarrollo de la enfermedad y la presencia o ausencia de algunos genes. Habiendo aplicado las pruebas estadísticas mencionadas en el capítulo tres para comprobar si la hipótesis sugerida es correcta o no, se puede concluir que con base en los datos analizados aquí, la hipótesis es cierta.

Para poder realizar un estudio de este tipo es necesario tomar en cuenta algunos factores como, los recursos disponibles, ya que de esto dependerán varios aspectos importantes dentro del estudio, por ejemplo, en el caso particular de éste, la muestra resultó ser muy pequeña debido principalmente a que en México no existen los aparatos necesarios para obtener la información requerida, por lo que la sangre tuvo que ser analizada en el extranjero y el costo de esto fue muy elevado.

Otro problema que disminuye la representatividad de los resultados fue la forma tan específica de escoger a las personas que iban a participar en el estudio, esto tiene que ver también con la parte económica, ya que al no contar con recursos suficientes se tuvo que escoger a los participantes de manera que si existía la sospecha de alguna asociación, ésta fuera más evidente utilizando una selección específica (personas con muchos años de diabetes y poca retinopatía, y viceversa, personas con pocos años de diabetes y mucha retinopatía). Problemas como éstos se van presentando a lo largo de los proyectos de investigación, lo que es un problema si lo que se quiere es minimizar esfuerzo y recursos, maximizando los resultados.

A pesar de los problemas arriba mencionados se obtuvieron resultados interesantes en los dos métodos, primero utilizando la prueba chi-cuadrada para el análisis de tablas de contingencia se puede decir que:

- La presencia del gen DR2 podría estar asociada con la presencia de la retinopatía diabética.
- La presencia del gen DR6, DR9 y B40 podría estar asociada con la ausencia de la retinopatía diabética.

Estos resultados fueron obtenidos con respecto a la población control, esto debido a que las diferencias significativas fueron encontradas al comparar la frecuencia génica de los controles, con la frecuencia génica de las personas diabéticas con retinopatía y sin ella. En el caso específico donde se consideró únicamente la frecuencia de las personas diabéticas con retinopatía contra la frecuencia de las que no tenían retinopatía (sin considerar a los controles), no arrojaron ningún resultado importante, por lo que no se puede decir nada acerca de si existe alguna asociación genética entre estos dos grupos.

Las conclusiones derivadas de utilizar el método de regresión logística se pueden observar en la tabla T.4.4, la cual se encuentra ordenada de manera descendente con respecto a la columna P(y=1), de ésta se puede decir lo siguiente: la participación de los genes A9 y DQ1 en los modelos es favorable, en el sentido de que si una persona que está sana, tiene dentro de su código genético los alelos anteriores entonces siempre tendrá una probabilidad muy grande (0.9921) de mantenerse sano que de desarrollar las enfermedades, otro dato importante que se puede ver en dicha tabla, es que la presencia del gen DR6 junto con la ausencia de A9, DR1, DR8 y DQ1, es la que arroja una probabilidad mayor de desarrollar diabetes (0.6659), mientras que, la ausencia del gen A9 junto con presencia de todos los demás, es la combinación que tiene la probabilidad mayor de desarrollar retinopatía (0.7264). Todas las posibles conclusiones se pueden observan en la tabla mencionada, siendo estas sólo las más representativas.

Derivado de que la muestra no cumple con las condiciones básicas de representatividad, las conclusiones aquí presentadas no se pueden tomar como definitivas, de ninguna manera se está diciendo con esto, que el estudio no fue de utilidad, sino que es necesario tal vez realizar un nuevo estudio con un tamaño de muestra mayor y tomando a los elementos de manera aleatoria, para confirmar los resultados aquí obtenidos.

# APÉNDICE 1

## ( 1 )

mlogit resp a2 a9 dr1 dr6 dr8 dq1

Iteration 0: Log Likelihood =-77.538675
Iteration 1: Log Likelihood =-61 5701
Iteration 2: Log Likelihood =-59.796018

Multinomial regression — Number of obs = 85
chi2(12) = 35.72
Prob > chi2 = 0.0004
Log Likelihood = -59.679093 — Pseudo R2 = 0.2303

| resp | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 2 | | | | | | |
| A2 | -1.008738 | .6253767 | -1.613 | 0.111 | -2.255704 | .2382274 |
| A9 | -2.459417 | .7872498 | -3.124 | 0.003 | -4.029148 | -.8896852 |
| DR1 | 2.149118 | 1.117685 | 1.923 | 0.059 | -.0794821 | 4.377718 |
| DR6 | 2.285161 | .8475616 | 2.696 | 0.009 | .5951715 | 3.975151 |
| DR8 | 1.202171 | .778757 | 1.544 | 0.127 | -.3506258 | 2.754969 |
| DQ1 | -2.148097 | .9534953 | -2.253 | 0.027 | -4.049312 | -.2468813 |
| _cons | - .0759495 | .5153167 | -0.147 | 0.883 | -1.103462 | .9515628 |
| 3 | | | | | | |
| A2 | -1.1225 | .9849691 | -1.140 | 0.258 | -3.086473 | .8414722 |
| A9 | -2.597372 | 1.258012 | -2.065 | 0.043 | -5.105777 | -.0889669 |
| DR1 | 3.042902 | 1.40064 | 2.173 | 0.033 | .2501046 | 5.8357 |
| DR6 | 3.214245 | 1.095 | 2.935 | 0.004 | 1.030878 | 5.397612 |
| DR8 | 2.000172 | 1.063729 | 1.880 | 0.064 | -.1208439 | 4.121187 |
| DQ1 | -1.81347 | 1.227336 | -1.478 | 0.144 | -4.260707 | .6337677 |
| _cons | -2.037476 | .9023502 | -2.258 | 0.027 | -3.836711 | -.238241 |

(Outcome resp==1 is the comparison group)

**( 2 )**

```
. mlogit resp a9 dr1 dr6 dr8 dq1

Iteration 0:   Log Likelihood    =-77.538675
Iteration 1:   Log Likelihood    =-62.922355
Iteration 2:   Log Likelihood    =-61.359678
Iteration 3:   Log Likelihood    =-61.264986
Iteration 4:   Log Likelihood    =-61.264192
Iteration 5:   Log Likelihood    =-61.264192

Multinomial regression                          Number of obs     =  85
                                                chi2(10)          =  32.55
                                                Prob > chi2       = 0.0003
Log Likelihood = -61.264192                     Pseudo R2         = 0.2099

------------------------------------------------------------------------------
        resp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
          A9 |  -2.358977   .7834002    -3.011   0.004    -3.920291   -.7976628
         DR1 |    2.27159   1.072927     2.117   0.038     .1332494    4.409931
         DR6 |   2.363444   .8547616     2.765   0.007     .6599065    4.066981
         DR8 |   1.374098   .7524414     1.826   0.072    -.1255151    2.873712
         DQ1 |  -2.089247   .9176747    -2.277   0.026    -3.91817    -.2603238
       _cons |   -.5442654   .4280118    -1.272   0.208    -1.397292    .3087608
-------------+----------------------------------------------------------------
3            |
          A9 |  -2.469552   1.243353    -1.986   0.051    -4.947552    .0084469
         DR1 |    3.16034   1.368914     2.309   0.024     .4320976    5.888582
         DR6 |   3.287807   1.091723     3.012   0.004     1.112006    5.463609
         DR8 |   2.188122   1.04415      2.096   0.040     .1071348    4.26911
         DQ1 |  -1.710072   1.195515    -1.430   0.157    -4.092731    .6725868
       _cons |  -2.554591   .8261768    -3.092   0.003    -4.201159   -.9080233
------------------------------------------------------------------------------
(Outcome resp==1 is the comparison group)
```

( 3 )

. mlogit resp a9 dr1 dr6 dr8

Iteration 0: Log Likelihood =-77.538675
Iteration 1: Log Likelihood =-65.546785
Iteration 2: Log Likelihood =-64.631977
Iteration 3: Log Likelihood = -64.60212
Iteration 4: Log Likelihood =-64.601971

Multinomial regression

Number of obs = 85
chi2(8) = 25.87
Prob > chi2 = 0.0011
Pseudo R2 = 0.1668

Log Likelihood = -64.601971

| resp | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 2 | | | | | | |
| A9 | -1.948415 | .7087214 | -2.749 | 0.007 | -3.36026 | -.5365691 |
| DR1 | .5974446 | .7092854 | 0.842 | 0.402 | -.8155243 | 2.010413 |
| DR6 | 1.587321 | .7015186 | 2.263 | 0.027 | .1898244 | 2.984818 |
| DR8 | 1.079003 | .6835417 | 1.579 | 0.119 | -.2826815 | 2.440688 |
| _cons | -.7972406 | .4127925 | -1.931 | 0.057 | -1.619565 | .0250843 |
| 3 | | | | | | |
| A9 | -2.046883 | 1.180762 | -1.734 | 0.087 | -4.399082 | .3053158 |
| DR1 | 1.810417 | .9809951 | 1.845 | 0.069 | -.1438259 | 3.764659 |
| DR6 | 2.608052 | .96874 | 2.692 | 0.009 | .6782225 | 4.537881 |
| DR8 | 1.913228 | .9979653 | 1.917 | 0.059 | -.0748209 | 3.901277 |
| _cons | -2.79974 | .8086528 | -3.462 | 0.001 | -4.410659 | -1.188821 |

(Outcome resp==1 is the comparison group)

# BIBLIOGRAFÍA


1. Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, John Wiley & Sons, Nueva York.

2. Asimov, I. (1973), *Introducción a la Ciencia*, Plaza y Janez, Barcelona.

3. Ayala, F. J. (1982), *Population and Evolutionary Genetics: A Primer*, The Benjamin/Cummings Publishing Company, Inc., California.

4. Bishop, Y. M. M. (1969), *Full Contingency Tables, Logits, and Split Contingency Tables*, Biometrics 25:383-400.

5. Bunge, M. (1972), *La investigación Científica*, Editorial Ariel, Barcelona.

6. Canavos, G. C. (1988), *Probabilidad y Estadística: Aplicaciones y Métodos*, 1ra. Edición, McGraw-Hill, México.

7. Chen, T., Fienberg, S. E. (1976), *The Analysis of Contingency Tables with Incompletely Classified Data*, Biometrics 32:133-144.

8. Everitt, B. S. (1980), *The Analysis of Contingency Tables*, 2da. Reimpresión, Chapman and Hall, Londres.

9. Fienberg, S. E. (1979), *The Use of Chi-squared Statistics for Categorical Data Problems*, J. R. Statis. Soc. B 41(1):54-64.

10. Gonick, L., Wheelis, M. (1991), *The Cartoon Guide to Genetics*, 1ra. Edición, HarperPerennial, Nueva York.

11. González-Villalpando, M. E., Arredondo-Pérez, B., González-Villalpando, C. (1994), *Retinopatía Diabética: Prevalencia y Severidad*, Rev Mex Oftalmol, 68(2):61-66.

12. Green, J. R., Mui Kui Chiew, Heng Chin Low, Woodrow, J. C. (1983), *The Association between HLA Antigens and the Presence of Certain Diseases*, Statistics in Medicine 2:79-85.

13. Grizzley, J. E., Starmer, C. F., Koch, G. G. (1969), *Analysis of Categorical Data by Linear Models*, Biometrics 25:489-504.

14. Grizzle, J. E., Williams, O. D. (1972), *Log Linear Models and Tests of Independence for Contingency Tables*, Biometrics 28:137-156.

15. Grupo de Reproducción y Genética AGN y Asociados (1996), *Ovulos y Espermas, Mensajeros del Código de la Vida*, Hospital Angeles 2(8).

16. Heredia-Ancona, B. (1991), *Introducción al Método Científico*, 4ta. Reimpresión, Editorial CECSA, México.

17. Hosmer, D. W. Jr., Lemeshow, S. (1989), *Applied Logistic Regresion*, John Wiley & Sons, Nueva York.

18. Lerman-Garber, I. (1994), *Atención Integral del Paciente Diabético*, McGraw-Hill, México.

19. Méndez Ramírez, I., Namihira Guerrero, D., Moreno Altamirano, L., Sosa de Martínez, C. (1993), *Protocolo de Investigación: Lineamientos para su elaboración y análisis*, 2da. Reimpresión, Editorial Trillas, México.

20. Nagel, E. (1973), *Introducción a la Lógica y al Método Científico*, Vol. 2, Amorrortu Editores, Buenos Aires.

21. Phillips, M., Del Rio, I., Quiroz, H. (1994), *Opportunities for Cost Reduction in Diabetic Retinopathy Treatment: Case Study from Mexico*, Bulletin of PAHO 28(1):50-60.

22. Popper, K. R. (1980), *La Lógica de la Investigación Científica*, Editorial Tecnos, Madrid.

23. Porta, J., McHugh, R. (1980), *Detection of HLA Haplotype Associations with Disease*, Tissue Antigens 15:337-345.

24. Rifkin, H. (1981), *Diabetes Mellitus*, Robert J. Brady Company, EUA.

25. Rivera, M. M. (1975), *Comprobación Científica de Hipótesis*, ANUIES, México.

26. Rull, J. A., Zorrilla, E., Jadzinsky, M. N., Santiago, J.V. (1992). *Diabetes Mellitus: Complicaciones Crónicas*, McGraw-Hill,

27. Simons, M. F., Tait, B. D. (1981), *Detection of Immune-Associated Genetic Markers of Human Disease*, Churchill Livingstone, Melbourne.

28. Weckmann, A. L., Vargas-Alarcón, G., López, M., González, N., De Leo, C., Castelán, N., Bordes, J., Alarcón-Segovia, D., Granados, J., Ramírez, E., Lisker, R. (1997), *Frequencies of HLA-A and HLA-B Alleles in Mexico City Mestizo Sample*, Am.J.Hum.Biol.,9:1-5.

29. Yuren-Camarena, M. T. (1975), *Leyes, Teorías y Modelos*, ANUIES, México.