ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS

ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS

LUIS GUILLERMO DÍAZ MONROY

MARIO ALFONSO MORALES RIVERA

Departamento de Estadística
Facultad de Ciencias

Universidad Nacional de Colombia
Sede Bogotá

Análisis estadístico de datos categóricos

Primera edición, 2009
Bogotá, Colombia
ISBN 978-958-719-186-8

Impresión:
Editorial Universidad Nacional de Colombia
direditorial@unal.edu.co
Bogotá, Colombia

Diseño de carátula: Andrea Kratzer M.

Catalogación en la publicación Universidad Nacional de Colombia

Díaz Monroy, Luis Guillermo, 1958-
Análisis estadístico de datos categóricos / Luis Guillermo Díaz Monroy, Mario Alfonso Morales Rivera. – Bogotá : Universidad Nacional de Colombia. Facultad de Ciencias, 2009
xxiv, 359 p.
ISBN : 978-958-719-186-8
1. Análisis de regresión logística 2. Tablas de contingencia 3. Modelos log-lineales 4. Análisis de correspondencias (Estadística) 5. Modelos lineales (Estadística) I. Morales Rivera, Mario Alfonso, 1965- II. Tít

CDD-21 519.536 / 2009

A Daniel, Camila, Diego y Pilar, mi única categoría.
Luis G. Díaz

A Nevis, mi ángel guardián.
Mario A. Morales

Contenido

Tablas

Figuras

Introducción

La distinción entre los llamados datos *cualitativos* y los denominados *cuantitativos* no siempre es clara, pues en algunos casos variables de tipo cuantitativo pueden considerarse como variables categóricas al dividir su rango de valores en intervalos o categorías, esto corresponde a una categorización de una variable cuantitativa. Un tratamiento recíproco puede considerarse para las variables cualitativas, es decir que pueden transformarse a variables cuantitativas, este procedimiento se muestra con el análisis de correspondencias. En estas notas se hace una revisión, bastante panorámica, sobre algunas metodologías estadísticas que coadyuvan al esclarecimiento e interpretación de la información contenida en datos categóricos.

El texto ha sido elaborado pensando en un lector que demande el uso de algunas herramientas estadísticas, útiles para el análisis de la información, principalmente de tipo categórico, producto de algún trabajo de investigación. No obstante que los primeros destinatarios son las personas que trabajen en torno a problemas de la salud y la biología, el material estadístico que se ofrece puede ser empleado por investigadores de otras disciplinas, pues basta cambiar el escenario de los ejemplos e ilustraciones, para hacer de este texto un instrumento de apoyo a varias disciplinas.

La primera parte contiene algunos conceptos generales junto con el tratamiento clásico de datos categóricos a través del análisis de tablas de contingencia, los cuales se desarrollan en los capítulos 1 y 2. Tratamientos alternativos a las tablas de contingencia se desarrollan en los capítulos 3 y 4 mediante el análisis de correspondencias y los modelos log-lineales. En el capítulo 5 se presenta el modelamiento con variable respuesta de tipo categórico el cual se hace a través de la regresión logística. El capítulo 6 contiene algunas de las técnicas de discriminación de uso más

frecuente. En el capítulo 7 se esquematizan algunos contrastes de tipo no paramétrico sobre estadísticas de localización. Se tratan, en el capítulo 8, algunas técnicas estadísticas para datos de conteo. Finalmente, en el capítulo 9, se desarrolla la técnica de emparejamiento de datos.

Para el desarrollo de los cálculos que las estimaciones y estadísticas requieren, se hace uso, principalmente, de los paquetes SAS y R. En cada capítulo se presenta la sintaxis pertinente para la ejecución de tales cómputos. Se debe advertir que existen otras herramientas computacionales igualmente útiles, tales como SPSS, BMDP, MINITAB, STATA, entre otras.

Agradecemos al Grupo de Investigación en "Estadística aplicada en la investigación experimental, la industria y la biotecnología", al Departamento de Estadística de la Universidad Nacional y al Departamento de Matemáticas y Estadística de la Universidad de Córdoba por posibilitar y permitir el ofrecimiento de estas notas.

Las notas se deben principalmente a la bibliografía que se anexa al final y a las preguntas, comentarios y sugerencias de nuestros colegas y de nuestros estudiantes. Este es un material que se puede mejorar en la medida que sea leído y cuestionado, por tanto agradecemos los comentarios y sugerencias que surjan de su estudio.

Luis Guillermo Díaz Monroy

Mario Alfonso Morales Rivera

Capítulo 1

Conceptos preliminares

1.1 Introducción

En este capítulo se presentan los aspectos fundamentales, a manera de elementos estadísticos básicos, para el desarrollo de los demás temas. Se revisa la naturaleza de los datos categóricos, el modelo probabilístico *binomial*, el de *Poisson* y la inferencia sobre una *proporción*.

1.2 Escala de medida

La escala de medida de una variable categórica es un elemento importante para la selección del análisis estadístico apropiado. Una selección inadecuada de la escala de medida puede conducir a una estrategia estadística inapropiada que arrojaría conclusiones erróneas acerca de la realidad contenida en los datos.

1.2.1 Dicotómicas

Son variables que tienen dos posibles respuestas, frecuentemente corresponden a la presencia o no de cierto atributo. Por ejemplo: ¿Desarrolló el sujeto la enfermedad? ¿En los últimos tres meses ha fumado alguna vez, o no? ¿Está afiliado actualmente al régimen contributivo de salud, o no?, etc.

Por ejemplo, el objetivo de un ensayo clínico para un nuevo medicamento contra la gripe es saber si los pacientes alivian sus dolencias. La tabla 1.1 contiene información sobre 124 pacientes, quienes recibieron tratamiento (medicamento) o un placebo (sin medicamento).

Tabla 1.1: Resultados respiratorios.

Tratamiento	Favorable	Desfavorable	Total
Placebo	16	48	64
Prueba	40	20	60

El grupo placebo consta de 64 pacientes; mientras que el grupo de prueba del medicamento contiene 60 pacientes.

1.2.2 Ordinal

En muchas ocasiones las variables categóricas representan más de dos posibles resultados, y a veces estos resultados poseen un orden propio. Tales variables tienen una escala de medida *ordinal*.

El estado de mejoría o progreso de un paciente se puede calificar como marcado (3), regular (2), ninguno (1). Este es el caso de un ensayo clínico en el que se investiga un tratamiento para la artritis reumatoidea. A hombres y mujeres les fue asignada una actividad (tratamiento) o un placebo (no actividad). Se midió el nivel de progreso o mejoría conseguido al final del ensayo; los datos están dispuestos en la tabla 1.2. Note que las variables categóricas pueden manejarse de diferentes for-

Tabla 1.2: Datos de artritis.

		Progreso			
Sexo	Tratamiento	Marcado	Regular	Ninguno	Total
Femenino	Actividad	16	5	6	27
Femenino	Placebo	6	7	19	32
Masculino	Actividad	5	2	7	14
Masculino	Placebo	1	0	10	11

Fuente: Stokes, Davis y Koch (1997: 218)

mas. Por ejemplo, en la tabla 1.2 se pueden fusionar las columnas Marcado y Regular para producir una variable dicotómica: "Progreso" frente

a "No progreso". Este tipo de agrupamiento se hace generalmente durante el análisis cuando hay interés por esta clase de respuestas o cuando se quiere obtener información adicional sobre los datos.

1.2.3 Conteos discretos

Corresponde a los casos en los que en lugar de registrar categorías, los resultados son números enteros. Por ejemplo, una investigación sobre enfermedades respiratorias en niños de diferentes zonas visitados dos veces determina si ellos mostraron síntomas de la enfermedad. La respuesta medida fue si los niños exhibieron síntomas en 0, 1 o 2 periodos. La tabla 1.3 contiene estos resultados.

Tabla 1.3: Niños con problemas respiratorios.

		Periodos			
Sexo	Tratamiento	0	1	2	Total
Femenino	Rural	45	64	71	180
Femenino	Urbana	80	104	116	300
Masculino	Rural	84	124	82	290
Masculino	Urbana	106	117	87	310

1.2.4 Nominal

Si se dispone de variables con más de dos categorías, a las cuales no se les atribuye un orden, se tiene una variable de tipo *nominal*. Por ejemplo, el tipo de sangre de una persona y la región geográfica donde nació; la tabla 1.4 muestra esta información. En este tipo de variables la relación que se puede establecer entre sus categorías es estrictamente de igualdad (o desigualdad).

1.3 Esquema de muestreo

Cuando el interés en el estudio es de tipo inferencial, los datos categóricos pueden proceder de diferentes esquemas de muestreo, sea este probabilístico o no. La naturaleza del muestreo determina los supuestos que

Tabla 1.4: Tipo de sangre por región de procedencia.

	Tipo de sangre				
Región	O	A	B	AB	Total
Norte	40	160	150	50	400
Centro	50	130	100	40	320
Sur	70	180	90	60	400
Total	160	470	340	150	**1.120**

pueden hacerse para desarrollar y aplicar un análisis estadístico determinado. Generalmente, los datos se ubican en uno de tres esquemas muestrales: datos históricos, datos experimentales y datos de encuestas.

Los datos históricos se refieren a estudios en los cuales los datos tienen una definición geográfica o circunstancial. Por ejemplo, la ocurrencia de una enfermedad infecciosa en una área determinada, los niños atendidos en un centro de salud, o el número de accidentes durante un periodo específico.

Los datos experimentales son extraídos de estudios que involucran la asignación aleatoria de tratamientos a un grupo de sujetos. Es el caso en el que a los sujetos se les administra una dosis entre varias dosis de un medicamento. En ciencias de la salud, los datos experimentales pueden incluir pacientes a quienes se les administra un placebo o un medicamento de acuerdo con sus condiciones médicas.

En estudios por encuestas, los individuos son seleccionados aleatoriamente desde una población objetivo. Por ejemplo, se seleccion una muestra de los usuarios de determinado medicamento para investigar algunos rasgos físicos de estos. El investigador puede seleccionar aleatoriamente una población de estudio y luego asignar aleatoriamente tratamientos a los individuos que resulten para el estudio.

La principal diferencia entre los tres esquemas de muestreo asociados a los ejemplos anteriores es el empleo de la aleatorización para obtenerlos. Los datos históricos no involucran aleatorización; en consecuencia, es difícil asumir que ellos representan determinada población. Los datos experimentales tienen una buena cobertura de la población, la cual está restringida por los tratamientos considerados en el protocolo del estudio; en el muestreo por encuestas, los datos tienen una muy buena cobertura de alguna población grande.

La unidad de aleatorización (en conexión con la unidad de observación) puede ser un sujeto o un conglomerado de sujetos. Además, la aleatorización puede aplicarse a sujetos, llamados estratos o bloques, con igual o desigual probabilidad. En muestreo por encuestas, esto puede conducir a diseños complejos, como el muestreo aleatorio estratificado, o un diseño de conglomerados por múltiples etapas. En estudios de diseño experimental, tales consideraciones llevan a estudios de medidas repetidas, datos longitudinales, entre otros.

1.4 Modelos de muestreo

El análisis de datos categóricos, o casi cualquier tipo de análisis estadístico, requiere supuestos acerca del mecanismo de aleatorización que genera los datos; esto es, el modelo probabilístico desde el cual se asume que son generados los datos. Se presentan a continuación las distribuciones de probabilidad de uso más frecuente en el análisis de datos categóricos.

1.4.1 Distribución de Poisson

La distribución de Poisson es un modelo probabilístico adecuado para evaluar la probabilidad de ocurrencia de un evento en intervalos de tiempo, longitud, superficie o volumen. Por ejemplo, el número de accidentes por semana en un tramo de carretera, el número de aves muertas por kilómetro cuadrado en una región. Si X es la variable aleatoria que cuenta el número de veces que un evento ocurre por intervalo (tiempo, longitud, superficie, etc), y si se tiene que el número promedio de eventos por intervalo es μ, las probabilidades de los posibles resultados $k = 0, 1, 2, \ldots$ se calculan mediante la expresión

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}, \quad \text{para } k = 0, 1, 2, \ldots \tag{1.1}$$

El término $k!$ es llamado el *factorial* de k y denota el producto de los k–primeros enteros, es decir, $k! = 1 \times 2 \times 3 \times \cdots \times k$, con $0! = 1$. El término $e^{-\mu}$ denota la *función exponencial*, algunas veces expresada como $\exp(-\mu)$; siendo $e \approx 2.7182$ el cual es la base de los *logaritmos*

naturales. Esto último significa que $e^a = b$ si y solo si $\ln(b) = a$. Por ejemplo $e^{0.7} = \exp(0.7) = 2.0$ corresponde a que $\ln(2.0) = 0.7$.

Suponga que el número de personas con infarto que acuden a una clínica tiene una tasa promedio de 2 por semana. Mediante el modelo de Poisson con $\mu = 2$, (i) la probabilidad de 0 infartos $(k = 0)$, y, (ii) de a lo más un infarto, en una semana cualquiera, por (1.1), es igual, respectivamente, a:

$$(i)\quad P(X = 0) = \frac{e^{-2}2^0}{0!} = e^{-2} = 0.1353$$
$$(ii)\quad P(X \leq 1) = \frac{e^{-2}2^0}{0!} + \frac{e^{-2}2^1}{1!} = 0.1353 + 0.2707 = 0.4060.$$

donde $P(X = 1) = e^{-2}(2^1)/1! = 0.2707$ es la probabilidad que se presente un infarto durante una semana. De (i) la probabilidad de que hayan infartos es: $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.1353 = 0.8647$

El valor esperado de la variable aleatoria X con distribución de Poisson es igual a su varianza; es decir,

$$\mathrm{E}(X) = \mathrm{var}(X) = \mu, \quad \sigma(X) = \sqrt{\mu} \tag{1.2}$$

Para el caso de los infartos por semana, si la tasa de ocurrencia de estos permanece constante de una semana a otra, entonces en un periodo largo el conteo de estos tendría una media de alrededor de 2 y una desviación estándar cercana a $\sqrt{2} = 1.41$.

De acuerdo con los parámetros dados en (1.2) se observa que la varianza se incrementa a medida que la media lo hace; los conteos tienden a variar más cuando el nivel de sus promedios es alto. Así, cuando el número de infartos por semana es 10, se observa que la variabilidad es más grande que cuando el número es igual a 2 por semana.

1.4.2 Distribución binomial

En el ejemplo anterior, el número de infartos fatales semanal es aleatorio. El número de infartos semanal, fatales o no, es también aleatorio. En muchas aplicaciones se tiene como fijo el número de veces que se presenta un fenómeno. En cada caso, el resultado es un evento A o no es el evento A (es A^c); entonces, se quiere registrar las veces que este fenómeno ocurre

con la característica determinada (A) en un número fijo de observaciones (n).

Un ejemplo es el caso de infarto fatal (A) o no fatal (A^c) y, en general, la ocurrencia o no de un evento; también es común hablar de *"éxito"* o *"fracaso"* para referirse al evento A o al A^c, respectivamente.

Una variable aleatoria X tiene distribución *binomial* si su función de probabilidad está dada por

$$P(X = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}, \text{para } k = 0, 1, 2, \ldots, n. \tag{1.3}$$

donde los dos parámetros n y $\pi = P(A)$ son tales que n es un entero no negativo y $0 \leq \pi \leq 1$. Se escribe $X \sim B(n, \pi)$. Para $n = 1$ la variable aleatoria se denomina de *Bernoulli*; es decir, que una variable aleatoria binomial es una suma de variables independientes tipo Bernoulli.

La cantidad $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ es el número de posibles arreglos de k-elementos (subconjuntos) que se pueden formar a partir de un conjunto que tiene n -elementos. Así, por ejemplo, el número de formas como se puede conformar un comité de 3 personas, escogidas de un grupo de 5 personas, es

$$\binom{5}{3} = \frac{5!}{3!2!} = 10$$

La figura 1.1 muestra tres casos especiales de esta distribución con el mismo valor $n = 10$ y $\pi = 0.3$, 0.5 y 0.8, respectivamente. La distribución exhibida en la figura 1.1a corresponde a $\pi = 0.3$, la cual es sesgada a la derecha; para $\pi = 0.5$, figura 1.1b, representa la distribución simétrica en torno a su media $\mu = n\pi = 5$. La distribución para $\pi = 0.8$ es sesgada hacia la izquierda, como se muestra en la figura 1.1c. Note que el sesgo se tiene para valores de π diferentes de 0.5. Para valores de n suficientemente grandes, y cualquier valor de π, la distribución tiende a ser simétrica en torno a su media.

Para ilustrar, sea X el número de infartos fatales registrados en $n = 10$ infartos. Suponga que por estudios anteriores o similares se tiene que la probabilidad π de que ocurra un infarto fatal es 0.2, es decir, $n = 10$ y $\pi = 0.2$. La probabilidad de (i) $X = 0$ infartos fatales (y por tanto $n - 0 = 10$ infartos no fatales), y, (ii) a lo más un infarto fatal, de acuerdo con la ecuación (1.3), es igual a:

(i) $P(X = 0) = \binom{n}{0}\pi^0(1-\pi)^{10} = \frac{10!}{0!10!}(0.2)^0(0.8)^{10} = 0.1074$

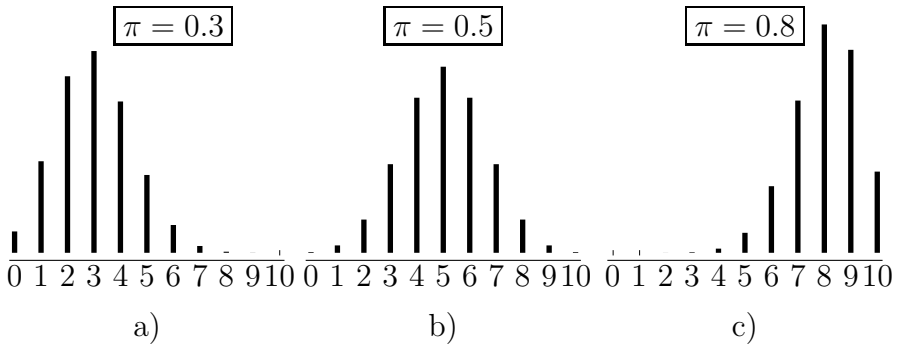

Figura 1.1: Distribución binomial.

(ii) $P(X \leq 1) = \binom{n}{0}\pi^0(1-\pi)^{10} + \binom{n}{1}\pi^1(1-\pi)^9$
$= 0.1074 + 0.2684 = 0.3758$

Así: (i) hay una probabilidad aproximanda al 11% de que se presenten cero infartos fatales, y (ii) una probabilidad de $26,87\%$ de que se presente a lo más un infarto (0 o 1), en muestras de tamaño 10 y en condiciones semejantes.

La distribución binomial para n ensayos independientes con probabilidad de "éxito" π tiene media y varianza

$$\mathrm{E}(X) = \mu = n\pi, \quad \mathrm{var}(X) = \sigma^2 = n\pi(1-\pi) \tag{1.4}$$

En el ejemplo considerado, la media es $\mu = n\pi = 10 \times 0.2 = 2$, y la varianza es $\sigma^2 = n\pi(1-\pi) = 10 \times 0.2 \times 0.8 = 1.6$.

En la tabla 1.5 se muestran las probabilidades asociadas al modelo de Poisson (valores de k mayores o iguales que 0) y al modelo binomial (valores de k entre 0 y 10).

1.4.3 Distribución multinomial

Es una generalización de la distribución binomial. A manera de ilustración suponga que un conjunto de n objetos puede clasificarse en k-clases distintas (*excluyentes y exhaustivas*). Por ejemplo, suponga que un grupo de n personas debe clasificarse de acuerdo con su grupo sanguíneo $\mathcal{A}$, $\mathcal{B}$, $\mathcal{AB}$ u $\mathcal{O}$. De esta forma, una persona se clasifica en una

Tabla 1.5: Distribución de Poisson y binomial con $\mu = 2.0$.

X	Poisson	Binomial
0	0.1353	0.1074
1	0.2707	0.2684
2	0.2707	0.3020
3	0.1804	0.2013
4	0.0902	0.0881
5	0.0361	0.0264
6	0.0120	0.0055
7	0.0034	0.0008
8	0.0009	0.0001
9	0.0002	0.0000
10	0.0000	0.0000
≥ 11	0.0000	-

sola de estas cuatro ($k = 4$) clases o modalidades respecto a su grupo sanguíneo.

En general, suponga que:

1. Se tiene una muestra de n elementos, cada uno de los cuales se asigna a una de las k clases.

2. El número de elementos que "caen" en la i-ésima clase es n_i, con $n_1 + n_2 + \cdots + n_k = n$.

3. Se define la variable aleatoria X_i $(i = 1, 2, \cdots, k)$, que "cuenta" el número de objetos ubicados en cada clase.

4. La probabilidad de que un objeto "caiga" en la i-ésima clase es p_i, con $p_1 + p_2 + \cdots + p_k = 1$.

5. La probabilidad de que se observen n_1 casos en la clase 1, n_2 casos en la clase 2, y así sucesivamente, que se observen n_k casos en la clase k, se calcula mediante la siguiente expresión:

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \quad (1.5)$$

A manera de ejemplo, suponga que, en un grupo humano determinado, la probabilidad de que una persona tenga sangre tipo $\mathcal{A}$ es 0.30; tipo $\mathcal{B}$ es 0.35; tipo $\mathcal{AB}$ es 0.20, y tipo $\mathcal{O}$ es 0.15. La probabilidad que en una muestra de 10 de estas personas haya 3 con sangre tipo $\mathcal{A}$, 4 con sangre tipo $\mathcal{B}$, 2 con sangre tipo $\mathcal{AB}$ y 1 con sangre tipo $\mathcal{O}$ es:

$$P(X_1 = 3,\ X_2 = 4,\ X_3 = 2, X_4 = 1) = \frac{10!}{3!4!2!1!}(0.30)^3(0.35)^4(0.20)^2(0.15)^1 = 0.03063.$$

El número $\frac{10!}{3!4!2!1!} = 12600$ es el número de formas como pueden distribuirse las 10 personas de manera que "caigan" en cada una de las 4 clases $\mathcal{A}$, $\mathcal{B}$, $\mathcal{AB}$ y $\mathcal{C}$, un número de personas de 3, 4, 2 y 1, respectivamente.

Un caso de distribución multinomial está asociado a tablas que cruzan dos variables categóricas (tablas de contingencia), pues las celdas de estas tablas corresponden a las modalidades o clases de una variable multinomial. En el ejemplo que se muestra en la tabla 1.1, la variable fila es el medicamento aplicado, con 2 modalidades; la variable columna es el estado de alivio del paciente, también con dos modalidades. Por tanto, las $2 \times 2 = 4$ celdas corresponden a las clases de una distribución multinomial.

1.4.4 Distribución hipergeométrica

Suponga que se tiene una población de tamaño N donde hay K elementos con el atributo A y $N - K$ sin el atributo A (A^c). La figura 1.2 ilustra esta población.

La variable hipergeométrica, X, cuenta el número de elementos con el atributo A que se obtienen en una muestra de tamaño n, de donde

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad \text{con } x = 0, 1, \ldots, \text{mín}\{K, n\} \tag{1.6}$$

es la probabilidad, de que en una muestra de n-objetos seleccionada de una población en la que hay K-objetos con el atributo A y $N - K$ sin él, se obtengan x-objetos con el atributo A, y, por tanto, $n - x$ sin este atributo.

A manera de ilustración, suponga que una caja contiene 10 manzanas, 2 de las cuales están dañadas; en este caso $N = 10$ y $K = 2$. Si

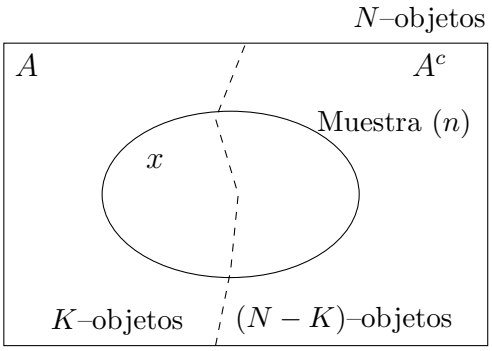

Figura 1.2: Esquematización de una distribución hipergeométrica.

seleccionan aleatoriamente $n = 3$ manzanas de la caja, la probabilidad que en la muestra resulten $X = 2$ manzanas dañadas es

$$P(X = 2) = \frac{\binom{2}{2}\binom{8}{1}}{\binom{10}{3}} = \frac{8}{120} = 0.0667.$$

Así, bajo las condiciones anteriores, se espera que en un muestreo repetitivo de muestras de tamaño 3 manzanas, aproximadamente el 7% de estas muestras contengan dos manzanas dañadas.

La distribución hipergeométrica es equivalente a la binomial cuando el muestreo se hace sin reemplazamiento.

1.5 Inferencia sobre una proporción

En algunas situaciones prácticas, los valores de los parámetros asociados con las distribuciones (como las anteriores) no se conocen. Una aproximación al conocimiento de estos parámetros es su *estimación* mediante los datos contenidos en una muestra de la población.

La *estimación* de un parámetro puede conducir a un valor específico, o a un intervalo dentro del cual se espera que, con cierta probabilidad o frecuencia, esté contenido el parámetro; estas estimaciones se conocen como estimación *puntual* y *por intervalo*, respectivamente.

También, sobre la distribución de una variable, y por consiguiente sobre los parámetros asociados a esta, se pueden hacer supuestos; estos se

conocen con el nombre de *hipótesis estadísticas*. En el problema de *verificar* o *contrastar* estos supuestos, se contempla una medida que da cuenta de la discrepancia entre el supuesto y lo observado a través de los datos.

En general, esta temática se conoce como *inferencia estadística*; que trata de conseguir y confrontar información sobre una población, respecto a un parámetro o distribución, a través de la que se dispone en una muestra. En otras palabras, la inferencia estadística es el procedimiento mediante el cual nos acercamos al conocimiento de una población con base en los datos disponibles en una muestra.

Los procedimientos de inferencia estadística pueden clasificarse en función de los supuestos requeridos y en función del tipo de información que utilicen. A continuación se señalan los métodos de inferencia estadística más empleados.

1. Respecto a los supuestos: *métodos paramétricos frente a no paramétricos*.

   Los métodos paramétricos suponen que los datos provienen de una distribución conocida cuyos parámetros se desea estimar. Los métodos no paramétricos no requieren el conocimiento de la distribución y solamente introducen hipótesis muy generales respecto a esta (continuidad, simetría, etc.), para estimar su forma o contrastar su estructura.

2. Respecto a la información utilizada: *métodos clásicos frente a bayesianos*.

   Los métodos clásicos suponen que los parámetros son cantidades fijas desconocidas y que la única información existente respecto a ellos está contenida en la muestra. Los métodos bayesianos consideran que los parámetros son variables aleatorias y permiten introducir información a priori sobre ellos a través de una distribución a priori.

   Los métodos clásicos ofrecen una respuesta simple a una mayoría de problemas de inferencia; tal respuesta es sustancialmente análoga a la obtenida con el enfoque bayesiano suponiendo poca información a priori. El enfoque clásico es más adecuado en la etapa de crítica del modelo, donde se pretende que los datos muestren por sí solos la información que contienen, con el menor número de restricciones posibles.

En estas notas el tratamiento bayesiano estará prácticamente ausente, no obstante, para el análisis estadístico de datos categóricos se cuenta con una amplia literatura sobre estas técnicas (Carlin & Louis 1998).

1.5.1 Estimación

Para un modelo probabilístico particular, se pueden sustituir los datos muestrales en la función de probabilidad y considerarla una función de los parámetros desconocidos. Por ejemplo, para $n = 10$, suponga un modelo binomial para el cual $X = 0$. De acuerdo con la fórmula de la binomial (1.3) con parámetro π, la probabilidad de este resultado es igual a

$$P(X = 0) = \frac{10!}{(0!)(10!)}\pi^0(1-\pi)^{10} = (1-\pi)^{10}.$$

Esta probabilidad está definida para todos los valores de π para los cuales $0 \leq \pi \leq 1$. La probabilidad anterior, evaluada en los datos muestrales, está en función del parámetro π, pues los datos son cantidades constantes; esta se conoce como una *función de verosimilitud*. La función de verosimilitud para $X = 0$ eventos A (0−éxitos) en 10 ensayos es $l(\pi) = (1-\pi)^{10}$. El objetivo es encontrar el valor de π en el cual la función de verosimilitud $l(\cdot)$ se maximiza. La figura 1.3 muestra diferentes funciones de verosimilitud para diferentes valores X (número de éxitos).

El *estimador máximo verosímil* del parámetro π es el valor de este para el cual la probabilidad de los datos observados toma el valor más grande dentro del rango de la variable. De otra manera, es el valor del parámetro en el cual la función de verosimilitud toma su máximo; es decir, se trata de encontrar la población, determinada por el parámetro, de donde más probablemente se obtuvo la muestra. En la figura 1.3 se insinúa el máximo de cada función de verosimilitud. Así, cuando en 10 ensayos se tienen $X = 0$ éxitos, el estimador máximo verosímil de π es igual a 0.0; para el caso de 10 ensayos con 4 éxitos, el estimador máximo verosímil de π es 0.4; finalmente, cuando en 10 ensayos ocurren 8-éxitos, el estimador máximo verosímil de π es 0.8. En general, el máximo se determina mediante el uso del cálculo, con lo cual se encuentra que el estimador de máxima verosimilitud para π es x/n. Se nota con $\widehat{\pi} = \frac{x}{n}$ para indicar que el estimador de π es $\widehat{\pi}$; aunque también es muy usada la letra p para señalar al estimador de π, es decir, $\widehat{\pi} = p$. De esta manera,

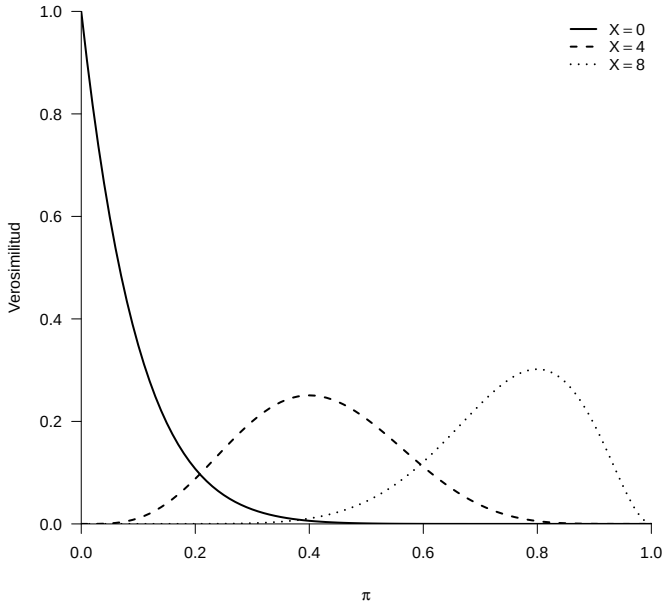

Figura 1.3: Función de verosimilitud para $X = 0$, 4 y 8-éxitos.

p es la *proporción* de éxitos en n-ensayos.

Si se considera la variable aleatoria X_i, $i = 1, 2, \ldots, n$, la cual toma el valor 1 si se presenta el "éxito" y 0 si se presenta un "fracaso", entonces la variable X, que cuenta el número de éxitos en los n ensayos, es igual a la suma de las X_i, es decir, $X = \sum_{i=1}^{n} X_i$. Así, el estimador de π es igual a la suma de la sucesión de ceros y unos dividida por el número de ensayos, esto es $p = \frac{\sum_{i=1}^{n} X_i}{n}$, con lo cual se muestra que p es una media aritmética de los X_i (ceros y unos). Para el caso $n = 10$ y $X = 4$, $p = (1+1+1+1+0+0+0+0+0+0)/10 = 0.4$. El orden de los ceros y de los unos, escritos en el numerador de la fracción, no necesariamente es igual al de los X_i; se debe leer que sólo ocurrieron 4 éxitos en los 10 ensayos.

1.5.2 Distribución muestral de una proporción

Suponga una población en que se observa la presencia o no de un atributo. Sea π la proporción desconocida de elementos con el atributo. La distribución muestral del estimador p, proporción observada en una muestra, se obtiene de la distribución binomial. Así, en una muestra aleatoria de tamaño n, de acuerdo con la fórmula (1.3), se tiene que

$$P\Big(p = \frac{k}{n}\Big) = P(X = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}, \text{ para } k = 0, 1, 2, \ldots, n.$$

Es decir, la probabilidad de que la proporción p sea igual a $\frac{k}{n}$ es equivalente a la probabilidad de obtener k "éxitos" en una muestra de tamaño n; que corresponde a la distribución binomial. Luego, por las ecuaciones (1.4), la media y la varianza de la distribución de p en el muestreo son, respectivamente,

$$\mathrm{E}(p) = \mathrm{E}(k/n) = \frac{n\pi}{n} = \pi \tag{1.7a}$$

y

$$\mathrm{var}(p) = \mathrm{var}(k/n) = \frac{1}{n^2}\,\mathrm{var}(k) = \frac{1}{n^2}n\pi(1-\pi) = \frac{\pi(1-\pi)}{n} \tag{1.7b}$$

Cuando el tamaño de muestra sea suficientemente grande, la distribución muestral de p será aproximadamente normal con la media y la varianza dadas en (1.7a) y (1.7b). Como se indica arriba, p es una media de variables dicotómicas, es decir, p se calcula por

$$p = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

donde cada X_i toma el valor de 1 si el elemento tiene el atributo (éxito), y 0 en otro caso. De esta manera, p tiene propiedades semejantes a la media muestral $\bar{X}$ en una población normal[1].

Note que la varianza de p, por (1.7b), implica el conocimiento de la proporción π, precisamente la que se pretende conocer a través de la muestra. De manera que la estimación de $\mathrm{var}(p)$ se obtiene mediante

$$\widehat{\mathrm{var}}(p) = \widehat{\sigma}^2(p) = \frac{p(1-p)}{n} \tag{1.8}$$

[1]Si $X_1, X_2, \cdots, X_n$, es una muestra aleatoria de una población normal, $N(\mu, \sigma^2)$, entonces, $\bar{X} \sim N\big(\mu, \frac{\sigma^2}{n}\big)$

Además se puede verificar que en tanto el tamaño de muestra (número de ensayos) aumenta, el *error estándar*, $\widehat{\sigma}(p)$, de p disminuye; es decir, la proporción muestral se acerca o aproxima al valor del parámetro π cuando n es grande.

Para una distribución de Poisson, por un procedimiento semejante, se encuentra que el estimador de máxima verosimilitud para μ, la tasa de eventos por intervalo, es $\widehat{\mu} = \bar{x}$; es decir, la media de las observaciones muestrales.

1.5.3 Intervalo de confianza para una proporción

En la sección anterior se señala que el estimador puntual de máxima verosimilitud para una proporción π es $\widehat{\pi} = p$, la proporción de eventos de interés (éxitos). Este se denomina *estimador puntual*, pues para cada conjunto de datos de una muestra el estimador p toma un valor específico. Como una alternativa o un complemento estaría el rango probable de valores para π; es decir un intervalo que permita establecer la incertidumbre existente en la estimación puntual. Esto se conoce como *intervalo de confianza*[2] para π.

Para una muestra de tamaño "grande", un intervalo del $(1-\alpha)$ de confianza para la proporción poblacional π se consigue mediante la expresión

$$p \mp Z_{1-\alpha/2}\widehat{\sigma}(p), \quad \text{con } \widehat{\sigma}(p) = \sqrt{p(1-p)/n} \tag{1.9}$$

donde $Z_{1-\alpha/2}$ es el percentil (sin signo) $1-\alpha/2$ de una distribución normal estándar, y $\widehat{\sigma}(p)$ es el *error estándar* estimado para p.

Los límites del intervalo de confianza dado en (1.9) significan que

$$P\Big[p - Z_{1-\alpha/2}\sqrt{p(1-p)/n} \leq \pi \leq p + Z_{1-\alpha/2}\sqrt{p(1-p)/n}\Big] = 1-\alpha.$$

Con una confiabilidad de 95% ($\alpha = 5\%$), un intervalo para π es

$$p \mp 1.96\sqrt{p(1-p)/n}$$

Por ejemplo, para el caso de los infartos con desenlace la muerte (fatales), suponga que en una muestra de $n = 242$ infartos registrados en los centros de salud de una zona determinada, 49 produjeron la muerte del paciente. De esta manera:

[2]La confianza corresponde a la probabilidad $1-\alpha$ (con $0 < \alpha < 1$) que el intervalo contenga al parámetro.

a) Un estimador puntual de π es $\widehat{\pi} = p = \frac{49}{242} = 0.20247$

b) Un estimador por intervalo de 95% de confianza para π es:

$$\begin{aligned} p \mp 1.96\sqrt{p(1-p)/n} &= 0.20247 \mp 1.96\sqrt{(0.20247)(0.79753)/242} \\ &= 0.20247 \mp 0.05063 \end{aligned}$$

que equivale a escribir el intervalo en la forma $[0.1518, 0.2531]$. Por tanto, se puede tener una confianza de 95% de que la proporción de infartos fatales, presentados en esta zona, está entre 0.1518 y 0.2531.

1.5.4 Contraste de hipótesis sobre una proporción

Para verificar la hipótesis nula de que el parámetro π es igual a cierta cantidad fija π_0; esto es $H_0 : \ \pi = \pi_0$, se utiliza la estadística

$$Z_0 = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \tag{1.10}$$

Esta estadística registra la "discrepancia" entre la muestra (sintetizada en el valor de p) y la hipótesis nula (el valor de π_0). Para tamaños de muestra grandes, la estadística Z_0 tiene distribución normal estándar[3].

La decisión de rechazar o no una hipótesis nula H_0, de acuerdo con la evidencia de los datos y el valor de verdad de H_0, se esquematiza en el cuadro siguiente Las decisiones ubicadas sobre la diagonal principal de

	$\overbrace{\qquad\qquad}^{H_0}$	
Decisión	*Cierta*	*Falsa*
No rechazo	DC	Error tipo II
Rechazo	Error tipo I	DC

la tabla se califican como decisiones correctas (DC), mientras que las que están fuera de tal diagonal se asumen como decisiones incorrectas. A la decisión de rechazar una hipótesis nula que es cierta se le denomina *error tipo I*; la decisión de no rechazar una hipótesis nula que es falsa

[3]De media $\mu = 0$ y varianza $\sigma^2 = 1$.

provoca que se incurra en el llamado *error tipo II*. El complemento del error tipo II corresponde al evento "rechazar la hipótesis nula cuando esta sea falsa"; la probabilidad de este evento se conoce como *potencia de la prueba*. La potencia de una prueba es su capacidad para detectar diferencias cuando realmente existen.

El valor máximo que se asume o tolera para el error tipo I se denomina *nivel de significancia de la prueba*; corresponde a la máxima probabilidad (riesgo) que se admite de estar rechazando una hipótesis que es cierta. El nivel de significancia se nota por

$$\alpha = P(Rechazar\ H_0|\ H_0\ es\ cierta).$$

De otra forma, el valor α significa la probabilidad de tener "discrepancias grandes" entre lo observado (muestra) y la hipótesis nula cuando esta es cierta; es decir, que se mantiene una desconfianza o "reserva" sobre la buena calidad y fidelidad de la muestra con relación a la población. Una notación semejante se tiene para la probabilidad del error tipo II: $\beta = P(No\ Rechazar\ H_0|\ H_0\ es\ falsa)$, luego $1 - \beta$ corresponde a la *potencia*; esta es la probabilidad de que la prueba detecte diferencias cuando realmente existan.

Con relación a la hipótesis nula sobre π, esta se rechaza para valores grandes, en valor absoluto (sin el signo), de Z_0. Valores grandes de Z_0, en valor absoluto, significan que los datos evidencian una alta discrepancia con la hipótesis nula, sea porque el valor de p es muy inferior al valor de p_0 o porque es muy superior a este.

Así, la hipótesis nula se rechaza si Z_0 se ubica en los extremos de una distribución normal estándar. En otras palabras, se rechaza H_0 si $Z_0 \geq Z_{\alpha/2}$ o si $Z_0 \leq -Z_{\alpha/2}$; en caso contrario, no se rechaza H_0.

El conjunto de valores de Z_0, en la distribución normal estándar, los cuales provocan el rechazo de la hipótesis nula, se denomina *región crítica*. La figura 1.4 ilustra la región crítica para este contraste (área sombreada). También suelen verificarse hipótesis de la forma $H_0 :\ \pi \geq p_0$ o $H_0 :\ \pi \leq p_0$, conocidas como *hipótesis unilaterales*, pues la región crítica corresponde a la cola derecha o la cola izquierda, en una distribución normal estándar, determinada por los valores Z_0 tales que $Z_0 \leq -Z_\alpha$ y $Z_0 \geq Z_\alpha$, respectivamente.

Una estrategia útil para la misma decisión, sobre la hipótesis nula, es el cálculo de la probabilidad de encontrar discrepancias mayores que o

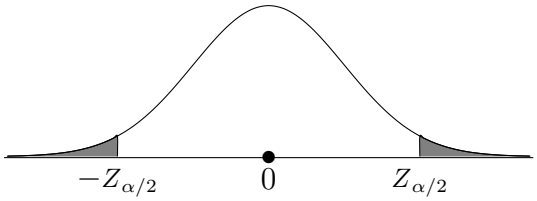

Figura 1.4: Región de rechazo para $H_0: \pi = p_0$.

iguales a la observada en la muestra. Esta probabilidad se conoce como el *valor p* o *nivel crítico p*. La mayoría de los paquetes estadísticos reportan el valor p asociado a una estadística de prueba evaluada en un conjunto de datos específico, de manera que la decisión de rechazar o no H_0 depende de la magnitud del valor p respecto al nivel crítico asumido; en consecuencia, si $p < \alpha$ se rechaza H_0, y en caso contrario, no se rechaza H_0.

Para tamaños de muestra demasiado pequeños la aproximación mediante la distribución normal puede resultar inadecuada. Una alternativa para estos casos es el empleo de la distribución binomial, es decir, la distribución exacta de la proporción muestral p.

Con la muestra en el caso de los infartos fatales, se quiere verificar la hipótesis "que la proporción de infartos fatales es igual a 20%", es decir, $H_0: \pi = 0.20$.

El valor de la estadística de prueba (1.10) en la muestra de las $n = 242$ personas es

$$\begin{aligned} Z_0 &= \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \\ &= \frac{0.20247 - 0.20}{\sqrt{\frac{(0.20)(0.80)}{242}}} \\ &= 0.09606. \end{aligned}$$

Con un valor de significancia $\alpha = 0.05$, como la región crítica es bilateral, el valor de $Z_{0.975} = 1.96$, luego, como $Z_0 = 0.09606 < 1.96$, no se rechaza la hipótesis nula; es decir la proporción de infartos fatales no difiere significativamente de 0.20. El valor p asociado a $Z_0 = 0.09606$ es 0.9232, esto es, $P(Z > 0.09606) = 0.4616$. La decisión es igual, puesto que el valor p es mayor que $\alpha/2 = 0.025$. En conclusión, los datos

muestrales corroboran la hipótesis nula.

1.6 Procesamiento de datos con R

En el apéndice B se hace una presentación sucinta de la sintaxis básica del paquete estadístico `R`, la cual aplicamos en esta sección.

Cálculo de probabilidades a partir de la distribución de Poisson con media $\mu = 2$.

```
# P(X=0)
dpois(0,lambda=2)

# P(X=1)
dpois(1,2)

#P(X<=1)
ppois(1,2)
```

Número de posibles arreglos de 3 elementos que se pueden formar a partir de un conjunto de 5, en notación de conteo, $\binom{5}{3}$ la cual se calcula con `R` mediante el código

```
choose(5,3)
```

Cálculo de probabilidades a partir de la distribución binomial con parámetros $n = 10$ y $\pi = 0.2$.

```
# P(X=0)
dbinom(0,size=10,prob=0.2)

# P(X=1)
dbinom(1,10,0.2)

# P(X<=1)=P(X=0)+P(X=1) (probabilidad acumulada hasta uno)
pbinom(1,10,0.2)
```

Intervalo de confianza para una proporción (sección 1.5.3).

```
prop.test(49, 242,correct=FALSE)
```

Prueba de la hipótesis $\pi = 0.20$, (sección 1.5.4)

```
prop.test(49, 242,p=0.2,correct=FALSE)
```

Si se requieren intervalos de confianza y pruebas de hipótesis para una proporción usando la distribución exacta del estadístico, se debe usar la función `binom.test( )`. Con el ejemplo anterior:

```
binom.test(49,242)
```

1.7 Ejercicios

1. La probabilidad de que una persona quede protegida al aplicarle determinada vacuna es $0,82$. Calcule la probabilidad de que una vez administrada a una muestra aleatoria de 15 pacientes:

   a) Ninguno sufra la enfermedad.

   b) Todos sufran la enfermedad.

   c) Al menos 12 no contraigan la enfermedad.

2. Suponga que a 100 personas se les aplica la vacuna a que se refiere el ejercicio 1, y 70 contraen la enfermedad.

   a) Estime, mediante un intervalo de confianza de 95%, la verdadera proporción de personas que quedan protegidas.

   b) ¿Existe evidencia en esta muestra de que la proporción de personas protegidas por la vacuna es 0.82? Justifique.

3. Para un volumen fijo, el número de células sanguíneas rojas es una variable aleatoria que se presenta con una frecuencia constante. Si el número promedio para un volumen dado es 9 células para personas normales, determine:

   a) La probabilidad de que el número de células rojas para una persona se encuentre dentro de una desviación estándar del valor promedio,

b) La probabilidad de que el número de células rojas para una persona se encuentre entre dos desviaciones estándar del valor promedio.

4. Cierta enfermedad tiene una probabilidad muy baja de ocurrir, $p = 0.00002$. Si 100 mil habitantes de una ciudad están expuestos, ¿cuál es la probabilidad de que menos de 2 personas presenten la enfermedad? Sugerencia: una variable aleatoria binomial con p pequeño y n grande puede modelarse mediante una Poisson con $\lambda = np$.

Capítulo 2

Tablas de contingencia

2.1 Introducción

En el presente capítulo se analizan datos que constituyen una muestra de una población clasificada respecto a dos o más variables categóricas. La tabla 2.1 contiene la clasificación de un grupo de 500 personas, con relación al estrato socioeconómico al que pertenecen y al concepto que tienen sobre el servicio de salud que reciben (Juez & Díez 1997, pág. 125).

Tabla 2.1: Opinión sobre el servicio de salud.

	Opinión			
Estrato	Bueno	Regular	Malo	Total
Bajo	75	35	40	150
Medio	60	70	50	180
Alto	20	30	40	90
Muy alto	15	25	40	80
Total	170	160	170	500

2.2 Tablas de contingencia

Una tabla como la 2.1 se conoce con el nombre de *tabla de contingencia*; en este caso, la variable fila (estrato) tiene 4 modalidades, mientras que la variable columna (opinión) tiene 3 modalidades. En general, una tabla de contingencia se asume como un arreglo bidimensional de f–filas por c–columnas ($f \times c$–celdas); en el ejemplo, $f = 4$ y $c = 3$. Las entradas en las celdas de la tabla son las frecuencias, o sea, el número de individuos de uno de los cuatro estratos con una de las tres opiniones; por ejemplo, en las modalidades medio y malo se ubican 50 personas. En general, se nota con n_{ij} a la frecuencia de la i–ésima modalidad de la variable fila y j–ésima de la variable columna; así, $n_{23} = 50$.

El total por fila o por columna está formado por las frecuencias marginales, y se notan por $n_{i.}$ (donde el punto señala que se suman columnas dentro de la fila i) y $n_{.j}$ (donde el punto señala que se suman filas dentro de la columna j), respectivamente. En la tabla 2.2, el número de personas que se ubican en el estrato alto es $n_{3.} = 90$, y es la frecuencia marginal para esta fila, mientras que la frecuencia marginal para la columna de opinión "regular" es $n_{.2} = 160$.

La suma de las frecuencias por celda es igual a la suma de las frecuencias marginales e igual al número total de individuos seleccionados y clasificados; se nota por N, en este caso $N = 500$.

La notación general, para una tabla de contingencia de f–filas y c–columnas, se muestra en la tabla 2.2 Donde

- La frecuencia o conteo de la i–ésima modalidad de la variable fila y la modalidad j–ésima de la variable columna se escribe como n_{ij}.

- El total de observaciones en la i–ésima modalidad de la variable fila se nota por $n_{i.}$, es decir,

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij} \tag{2.1}$$

- El total de observaciones en la j–ésima modalidad de la variable

Tabla 2.2: Tabla de contingencia.

Filas	Columnas						
	1	2	$\cdots$	j	$\cdots$	c	Total $(n_{i.})$
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
f	n_{f1}	n_{f2}	$\cdots$	n_{fj}	$\cdots$	n_{fc}	$n_{f.}$
Total $(n_{.j})$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	$n_{..} = N$

columna se nota por $n_{.j}$; es decir:

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{fj} = \sum_{i=1}^{f} n_{ij} \tag{2.2}$$

- El número total de observaciones (individuos) en la muestra se escribe con N, y es igual a la suma de los márgenes fila o columna (o a la suma de todas las celdas i, j); es decir:

$$N = \sum_{i=1}^{f} \sum_{j=1}^{c} n_{ij} \tag{2.3}$$

Las frecuencias pueden ser transformadas en proporciones o porcentajes. El primer porcentaje, para una tabla de contingencia, se obtiene de dividir cada frecuencia n_{ij} por el número total de observaciones N; este porcentaje se escribe como f_{ij}, es decir,

$$f_{ij} = \frac{n_{ij}}{N} \times 100 \tag{2.4}$$

la cantidad f_{ij} corresponde a la proporción o porcentaje de individuos que tienen los atributos i y j; o sea, en las dos variables.

El segundo porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal fila $n_{i.}$, esto es:

$$f_{j|i} = \frac{n_{ij}}{n_{i.}} \times 100 \tag{2.5}$$

La cantidad $f_{j|i}$ es la proporción de individuos de cada celda, respecto al total de la fila i. La expresión $j|i$ (que se lee: "j dado i") significa "estar en la columna j, a condición de estar en la fila i", es decir, se deja fija la fila i y se recorren sus columnas. Estas frecuencias corresponden al *perfil fila*. A continuación se muestra el perfil de la opinión sobre las personas clasificadas en estrato socioeconómico *alto*:

Estr. Alto:	Bueno	Regular	Malo	Total
n_{3j} :	20	30	40	90
$f_{j\|3} = \frac{n_{3j}}{n_{3.}} \times 100$:	$f_{1\|3} = 22.22$	$f_{2\|3} = 33.33$	$f_{3\|3} = 44.45$	100.00

El tercer porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal columna $n_{.j}$.; esto es:

$$f_{i|j} = \frac{n_{ij}}{n_{.j}} \times 100 \tag{2.6}$$

La cantidad $f_{i|j}$ es la proporción de individuos de cada celda, respecto al total de la columna j. La expresión $i|j$ (se lee: "i dado j") significa "estar en la fila i, a condición de estar en la columna j", es decir, se deja fija la columna j y se recorren sus filas. Estas frecuencias corresponden al *perfil columna*.

En consecuencia, se pueden obtener tres tipos de tablas adicionales: la primera hace referencia al porcentaje de cada celda con relación al número total de individuos N; la segunda, al porcentaje de cada celda respecto al total de la respectiva fila (perfil fila); y la tercera, al porcentaje de cada celda con relación al total de la respectiva columna (perfil columna). La tabla 2.3 contiene las cuatro tablas descritas anteriormente para los datos de la tabla 2.1 Los datos que se presentan en cualquiera de las últimas formas, proceden originalmente de conteos o frecuencias de variables categóricas más que de mediciones continuas. Es más, las variables continuas pueden transformarse en variables categóricas mediante la definición de intervalos sobre su escala; así, ejemplo, la temperatura de una persona, que es una variable continua, se puede considerar con arreglo a estas tres categorías: baja o hipotermia (menos de 36^oC), media o normal (entre 36^oC y 38^oC) y alta o hipertermia (mayor de 38^oC).

Cada una de estas tablas se puede representar gráficamente mediante un histograma de frecuencias. A continuación se presentan las gráficas (figuras (2.1a) y (2.1b)) para el porcentaje general: la primera muestra la

Tabla 2.3: Tabla de contingencia completa (de la tabla 2.1).

Estrato	*Opinión sobre salud*			
Frecuencia Porcentaje Porc. Fila Porc. Columna	Bueno	Regular	Malo	Total
Bajo	75	35	40	150
	15.00	7.00	8.00	30.00
	50.00	23.33	26.67	
	44.12	21.88	23.53	
Medio	60	70	50	180
	12.00	14.00	10.00	36.00
	33.33	38.89	27.78	
	35.29	43.75	29.4	
Alto	20	30	40	90
	4.00	6.00	8.00	18.00
	22.22	33.33	44.44	
	11.76	18.75	23.53	
Muy alto	15	25	40	80
	3.00	5.00	8.00	16.00
	18.75	31.25	50.00	
	8.82	15.63	23.53	
Total	170	160	170	500
	34.00	32.00	34.00	100.00

opinión en relación con el estrato socioeconómico; la segunda muestra el estrato socioeconómico frente a la opinión. En cada una de las gráficas, las modalidades (barras) conservan el orden que tienen en la tabla 2.3; es decir, izquierda–derecha y arriba–abajo, respectivamente.

La figura 2.2 muestra los perfiles condicionados a cada uno de los cuatro estratos. Se advierte una alta diferencia respecto a la forma y el tamaño de los histogramas; esto permite aventurar la hipótesis de asociación entre las variables fila y columna, porque al cambiar de una modalidad a otra del estrato (figuras (2.1a) y (2.2)), se producen modificaciones considerables en los tipos de opinión sobre el sistema de salud. Si no hubiera asociación, al desplazarse de una modalidad a otra de una variable, no se observaría algún patrón de cambio en las modalidades de la otra variable. Esta situación es contraria a la que se aprecia en las gráficas anteriores; por ejemplo, note que en la medida que el estrato aumenta, el concepto

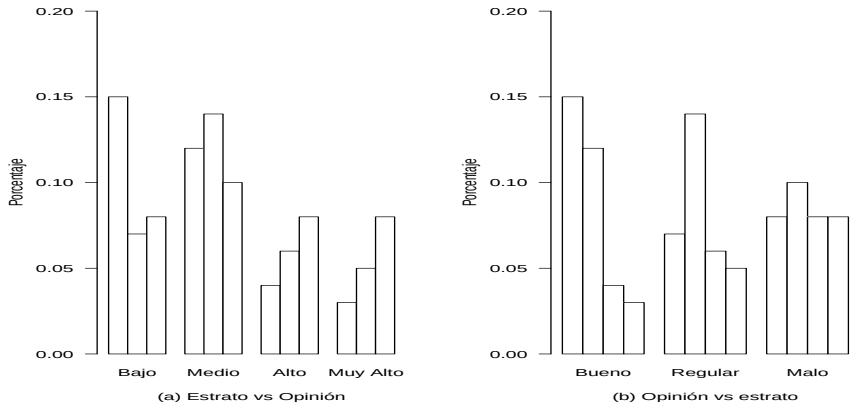

Figura 2.1: Distribución de frecuencias.

"favorable" sobre las instituciones de salud decrece. Esta información advierte acerca de la existencia de una posible asociación negativa entre estas variables. Una comparación entre las formas de los diferentes histogramas advierte sobre la posible asociación entre las variables; así histogramas isomorfos sugieren no asociación (independencia) entre el par de variables categóricas.

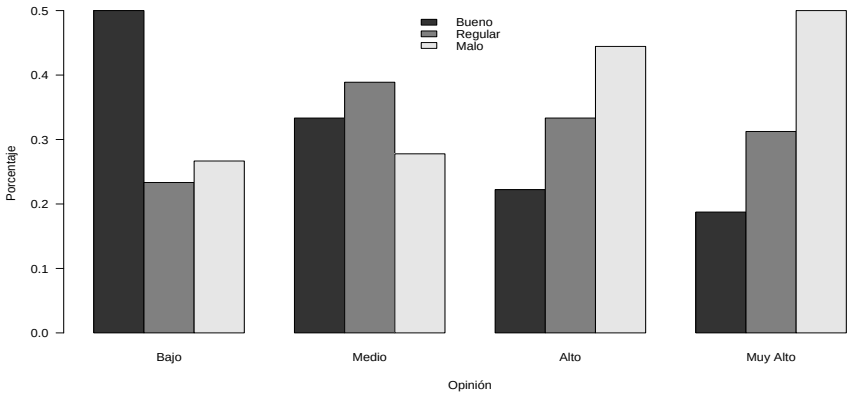

Figura 2.2: Perfiles fila de la opinión por estrato.

2.3 Modelos probabilísticos

La validez de la inferencia estadística se apoya en el cumplimiento de los supuestos probabilísticos en una población. El análisis de los datos de una tabla de contingencia se basa en los siguientes tres tipos de "mecanismos" aleatorios, los cuales, se supone, generan la estructura de los datos contenidos en una tabla de contingencia.

2.3.1 Modelo de clasificación fija

Es un caso no muy frecuente en investigación social. La fijación por muestreo de las probabilidades marginales de las variables fila y columna caracteriza este modelo.

Un ejemplo de este modelo de clasificación fija es el siguiente: un cuestionario de evaluación, sobre uno de los funcionarios en una institución de salud, fue aplicado a 30 de los usuarios atendidos por el funcionario: 15 hombres y 15 mujeres. El funcionario afirma que el perfil de respuestas de los hombres es distinto al de las mujeres. Para verificar esta afirmación, se permite que, con base en el examen de las respuestas, clasifique los formularios en grupos de 15, según crea que quien lo respondió se trata de un hombre o una mujer, sin que conozca el género real de las personas. La tabla 2.4 muestra los datos. En este caso los totales

Tabla 2.4: Evaluación de un funcionario.

	Clasificación del funcionario		
Género	Hombre	Mujer	Total
Hombre	6	9	15
Mujer	9	6	15
Total	15	15	30

fila ($n_{1.} = 15$ y $n_{2.} = 15$) y de columna ($n_{.1} = 15$ y $n_{.2} = 15$) se fijan por muestreo para contrastar la afirmación del funcionario; en consecuencia, el valor obtenido en cualquiera de las cuatro casillas determina el valor de las demás. La distribución de probabilidades para los datos de las celdas es la hipergeométrica (sección 1.4.4). En esta situación de muestreo, Fisher (1928-1958) propuso que la probabilidad de frecuencia asociada a una celda de una tabla 2×2, bajo la hipótesis nula de no asociación

entre el género de la persona que respondió el cuestionario y el género asignado por el funcionario, sigue una distribución hipergeométrica, la cual se calcula por

$$P_I = P(n_{11}|n_{..}, n_{1.}, n_{.1}) = \frac{n_{1.}!n_{.1}!n_{2.}!n_{.2}!}{n_{..}!n_{11}!n_{12}!n_{21}!n_{22}!} \tag{2.7}$$

Las demás probabilidades se calculan de manera semejante, pero una vez que se tiene el valor de la frecuencia de una celda o el valor de su probabilidad, los demás valores quedan determinados. Estas probabilidades están involucradas en la prueba *exacta de Fisher*, la cual se aborda en el estudio de tablas 2×2 para verificar la hipótesis de no asociación entre las variables fila y columna de estas tablas.

2.3.2 Modelo de homogeneidad

El modelo que genera estos datos se caracteriza porque el número de individuos muestreados en una de dos subpoblaciones (definidas por las dos variables categóricas) se clasifica en una de la variables; es decir, se tienen probabilidades marginales fijas para una variable y aleatorias para la otra.

Un ejemplo de este tipo de modelo es el siguiente: a un grupo de estudiantes universitarios se les aplica un cuestionario para conocer su concepto sobre la legalización del aborto en el país. De un total de 300 hombres entrevistados, 110 se manifestaron a favor, mientras que 150 de 200 mujeres rechazaron la iniciativa. De aquí se nota que los totales fila ($n_{1.} = 300$ y $n_{2.} = 200$) son valores fijados por muestreo, mientras que los totales columna son valores aleatorios. La tabla 2.5 contiene estos datos. Una consecuencia es que las probabilidades de la tabla de contingencia

Tabla 2.5: Concepto sobre el aborto.

	Concepto		
Género	A favor	En contra	Total
Hombre	110	190	300
Mujer	50	150	200
Total	160	340	500

son probabilidades condicionales. La probabilidad de que una persona

esté a favor, dado que es hombre, es $P(F|H) = \frac{110}{300} = 0.37$; mientras que la probabilidad de que esté en contra es $P(C|H) = \frac{190}{300} = 0.63$. De forma análoga, la probabilidad de que una persona esté a favor, dado que es de sexo femenino, es $P(F|M) = \frac{50}{200} = 0.25$ y la probablildad de que esté en contra es $P(C|M) = \frac{150}{200} = 0.75$.

Se supone que para cada categoría de la variable género, el número de individuos que está en una de las dos categorías, a favor o en contra, constituye una variable aleatoria binomial, de forma que en este contexto las tablas 2×2 se describen conforme a un muestreo de una *binomial doble*. La probabilidad exacta de una tabla de contingencia observada en el modelo de *homogeneidad* viene dada por

$$P_{II} = P_I \times P(n_{.1}|\pi, n_{..}) = P_I \frac{n_{..}}{(a+c)!(b+d)!} \pi^{(a+c)}(1-\pi)^{(b+d)} \quad (2.8)$$

Donde P_I es la probabilidad calculada con (2.7) y $P(n_{.1}|\pi, n_{..})$ es la probabilidad binomial del marginal aleatorio $n_{.1}$. Note que la probabilidad para un modelo de homogeneidad es menor que la de un modelo de clasificación fija ($P_{II} < P_I$).

Este modelo de muestreo es característica tanto de estudios *prospectivos* como los de *cohorte o de seguimiento*, donde las muestras se constituyen sobre la base de presencia o ausencia de alguna característica en una variable de clasificación (en el caso anterior la variable género) y se registra una variable respuesta, como de estudios *retrospectivos*, donde las muestras se constituyen sobre la base de la presencia o ausencia de alguna característica de la variable respuesta y después se determinan retrospectivamente las frecuencias sobre una variable de clasificación.

2.3.3 Modelo de independencia

Se caracteriza porque una muestra de individuos en una población se clasifica simultáneamente en dos variables, situación que refleja la existencia de probabilidades marginales aleatorias para ambas variables.

A manera de ilustración, suponga que una muestra de 250 estudiantes jóvenes se le indaga sobre el consumo o no de drogas y sus prácticas o no de relaciones bisexuales. Los datos resultantes conforman la tabla de contingencia 2.6. La única restricción que impone este modelo es que la suma de las probabilidades sea igual a 1.0, en consecuencia, existe una población única de donde procede la muestra, clasificada de acuerdo con

Tabla 2.6: Drogas vs. prácticas bisexuales.

	Bisexualidad		
Consumo de drogas	No	Sí	Total
No	165	15	180
Sí	52	18	70
Total	203	47	250

las dos variables categóricas (fila y columna), que representan las variables aleatorias. Ambos conjuntos de frecuencias marginales se considera que siguen varias distribuciones binomiales y únicamente el valor del tamaño muestral se conoce de antemano.

Para el caso de tablas 2×2, la probabilidad exacta para una tabla observada en este modelo es

$$\begin{aligned} P_{III} &= P_I \times P(n_{1.}|\pi', n_{..})P(n_{1.}|\pi, n_{..}) \\ &= P_{II}\frac{N}{(a+b)!(c+d)!}\pi'^{(a+b)}(1-\pi')^{(c+d)} \end{aligned} \tag{2.9}$$

Donde P_I es la probabilidad hipergeométrica de la tabla asociada a la ecuación (2.7) y los términos siguientes son probabilidades binomiales de las dos marginales observadas; π y π' son las probabilidades de respuesta asociadas.

Una alternativa a este modelo se produce cuando el tamaño muestral no es prefijado, sino que se considera aleatorio. En tales casos, la distribución más apropiada es la de Poisson.

2.4 Independencia de la clasificación

Una vez que se han dispuesto los datos en una tabla de contingencia, la pregunta más importante, en general, es si las variables cualitativas que conforman la tabla de contingencia son *independientes* o *no*. Veamos qué encierra la independencia: del ejemplo anterior es claro que si la forma como responden las personas encuestadas es independiente del estrato socioeconómico al que pertenecen, entonces la proporción (o porcentaje) de los que responden en las categorías Bueno, Regular y Malo es la misma para cada uno de los estratos Bajo, Medio, Alto y Muy alto; es decir,

que los respectivos histogramas deben ser geométricamente congruentes. Si estas proporciones difieren, esto puede ser asociado más a algunas de las modalidades de la otra variable; por supuesto que puede tenerse algún grado de diferencia en las proporciones debido a la aleatoriedad de la muestra y a otras causas no identificables. Lo que se necesita es saber si se puede o no asegurar si las diferencias observadas entre las dos proporciones es demasiado grande para que sea atribuida a tales causas. Por esto, en la sección 2.4.1 se muestra una prueba estadística de esta afirmación.

Suponga que en la población de donde se extrae la muestra, la probabilidad de que una observación pertenezca a la i–ésima modalidad de la variable fila y a la j–ésima modalidad de la variable columna es π_{ij}; en consecuencia, el número de observaciones que se espera "caigan" en la celda (i, j) (F_{ij}) de la tabla de acuerdo con las N observaciones muestreadas es dada por

$$F_{ij} = N\pi_{ij} \tag{2.10}$$

Ahora, $\pi_{i.}$ representa la probabilidad, en la población, de que una observación pertenezca a la i–ésima categoría de la variable fila; de manera semejante, $\pi_{.j}$ representa la probabilidad correspondiente para la j–ésima categoría de la variable columna. Entonces, la independencia entre las dos variables, en la población, se garantiza por el cumplimiento de la siguiente igualdad (Juez & Díez 1997, 11)

$$\pi_{ij} = \pi_{i.}\pi_{.j} \tag{2.11}$$

Es decir, que las dos variables son independientes si y solo si la probabilidad de que una observación "caiga" en la celda (i, j) es igual al producto de las respectivas probabilidades marginales. Desde el punto de vista de la frecuencia esperada en la tabla de contingencia (ecuación 2.10), por la independencia (ecuación 2.11), al reemplazar π_{ij} se tiene:

$$F_{ij} = N\pi_{i.}\pi_{.j} \tag{2.12}$$

Como los valores de π_{ij}, $\pi_{i.}$ y $\pi_{.j}$ son desconocidos en la población sobre la cual precisamente se indaga en un estudio, la verificación de la igualdad anterior no es fácil de realizar, razón por la que "debemos" conformarnos con la verificación a través de los datos muestrales. De esta manera, los valores de π_{ij}, $\pi_{i.}$ y $\pi_{.j}$ deben estimarse a través de la información contenida en la muestra. Desde los datos muestrales disponibles,

los respectivos estimadores para cada una de estas cantidades desconocidas, como se mencionó en la sección (1.5.1), son:

$$\begin{aligned}\widehat{\pi}_{ij} &= p_{ij} = \frac{n_{ij}}{N} \\ \widehat{\pi}_{i.} &= p_{i.} = \frac{n_{i.}}{N} \\ \widehat{\pi}_{.j} &= p_{.j} = \frac{n_{.j}}{N}\end{aligned} \tag{2.13}$$

donde los $\widehat{\pi}$ representan la probabilidad estimada mediante los datos muestrales. Así, la frecuencia esperada estimada, notada por E_{ij}, de acuerdo con las ecuaciones (2.12) y (2.13), es igual a

$$\begin{aligned}E_{ij} &= N\widehat{\pi}_{i.}\widehat{\pi}_{.j} = N\frac{n_{i.}}{N}\frac{n_{.j}}{N} \\ &= \frac{n_{i.}n_{.j}}{N}\end{aligned} \tag{2.14}$$

Es decir, que bajo independencia, la frecuencia esperada en cada celda es igual al producto de la frecuencia de las márgenes dividido por el número total de individuos de la tabla.

Cuando las dos variables sean independientes, la frecuencia estimada usando la ecuación (2.14) diferiría de las frecuencias observadas n_{ij} en cantidades muy pequeñas y atribuibles a cambios en las variables únicamente. Sin embargo, si las dos variables no son independientes se esperaría que se presentasen diferencias grandes entre las frecuencias observadas y las esperadas. Esta es la base de una medida que registra la independencia o no entre dos variables de tipo categórico, la cual se describe a continuación.

2.4.1 Prueba ji–cuadrado

Aunque existe una variedad amplia de estadísticas para contrastar la hipótesis de independencia entre variables categóricas, se presenta en esta sección la estadística de uso y abuso más frecuente: la ji–cuadrado. Para verificar la independencia entre dos variables se requiere averiguar por la veracidad de la hipótesis que de este supuesto se deriva, es decir por

$$\pi_{ij} = \pi_{i.}\pi_{.j} \tag{2.15}$$

En general, esta hipótesis se referirá como la *hipótesis nula*. Como se anotó, la verificación (no la prueba) de esta hipótesis debería basarse en la diferencia entre los valores estimados de las frecuencias que se esperan bajo la hipótesis nula (independencia) E_{ij} y las frecuencias observadas en la muestra n_{ij}. Tal prueba, sugerida por Pearson (1904), emplea la estadística χ^2 (ji–cuadrado), dada por

$$\chi_0^2 = N \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(p_{ij} - \widehat{\pi}_{i.}\widehat{\pi}_{.j})^2}{\widehat{\pi}_{i.}\widehat{\pi}_{.j}} \tag{2.15a}$$

o también por

$$\chi_0^2 = \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{2.15b}$$

La primera ecuación (2.15a) está formulada en términos de probabilidades empíricas y esperadas, y la segunda (2.15b) en términos de frecuencias observadas y esperadas.

La doble suma indica que se deben sumar todas las celdas, lo cual se consigue "desplazándose" por filas o por columnas. Se observa que la magnitud de esta estadística depende de los valores que asuman las diferencias $(n_{ij} - E_{ij})^2$. Note que la fórmula tiene implicada la distancia al cuadrado (euclidiana) entre las frecuencias observadas y las esperadas; de tal manera, la fórmula registra qué tan "distante" está lo observado de lo esperado bajo la hipótesis de independencia . Si las dos variables son independientes, estas diferencias deberían ser menores que en el caso contrario (dependencia); en consecuencia, χ_0^2 será más pequeña cuando la hipótesis nula es cierta que cuando es falsa. Por tanto, se necesita un método mediante el cual se pueda o no rechazar la hipótesis nula. Tal procedimiento se basa en la distribución de probabilidades de la estadística χ^2 bajo el supuesto de que la hipótesis de independencia es cierta. Para cada valor que tome la estadística χ^2 se tiene una probabilidad de que haya valores mayores o iguales (o menores) que tal valor, de manera que la distribución de probabilidades de χ^2 corresponde a las probabilidades asociadas a los diferentes valores de χ^2. Por construcción, la estadística ji–cuadrado toma valores positivos, pues es una suma de cantidades elevadas al cuadrado. La figura 2.3 muestra la distribución de probabilidades asociada a la estadística ji–cuadrado. La decisión de rechazar o no la hipótesis nula de independencia entre las variables se basa en la probabilidad de los valores obtenidos para χ_0^2; pues valores

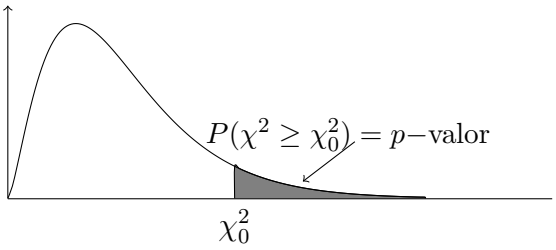

Figura 2.3: Región de rechazo para $H_0 : \pi = p_0$.

con baja probabilidad (de ser mayores o iguales que este) permiten el rechazo de la hipótesis nula o, en caso contrario, su no rechazo.

Como se nota en la gráfica, valores sobre el eje horizontal (χ^2) con baja probabilidad a la izquierda tienen probabilidades altas a la derecha; y, recíprocamente, valores con alta probabilidad a la izquierda tienen baja probabilidad a la derecha. Como se explicó, valores de χ^2 con baja probabilidad (a la derecha) provocan el rechazo de la hipótesis nula; pues esto dice que la probabilidad de que no existan discrepancias entre los valores observados y los esperados, bajo la hipótesis de independencia, es baja, pues como se anotó, la magnitud de la estadística χ_0^2 depende de los valores que asuman las diferencias $(n_{ij} - E_{ij})^2$: si las dos variables son independientes, estas diferencias deberían ser menores que en el caso contrario (dependencia).

Este es el procedimiento usual para derivar la llamada *significancia de la prueba*. Se entiende por significancia la probabilidad de rechazar la hipótesis nula (no asociación) cuando esta es verdadera (asociación). En general, se consideran valores de significancia "bajos" a valores de probabilidad menores que 0.05 (5%); valores por debajo de 0.01 (1%) se consideran como extremadamente bajos, estos son referidos como *niveles de significancia* de la prueba estadística. Una de las inquietudes que encuentran los usuarios de estos métodos estadísticos, es el umbral a partir del cual las conclusiones derivadas adquieren un carácter objetivo y aceptable. Es decir, ¿qué tan grande debe ser el nivel de significancia para ser considerado como bajo, aceptable o alto? Esta respuesta no la tienen los estadísticos, pues depende del ámbito de investigación, de los intereses que sobre el objeto en cuestión se planteen, e incluso, del grado de ignorancia admitido sobre el objeto.

Podemos encontrar valores de significancia bajos que en ciertos contex-

tos advierten sobre la no asociación, pero que en otros contextos son suficientes para emprender una búsqueda, "una luz en donde antes sólo había oscuridad". Así mismo, podemos encontrar valores con los que se evidencia una asociación significativa y que, no obstante, no nos deben forzar a abandonar el escepticismo encerrado en ese pequeño valor de incertidumbre, "una oscuridad en donde parece haber luz".

2.4.2 Distribución ji–cuadrado

La distribución ji–cuadrado se puede definir como la suma de los cuadrados de distribuciones normales estándar independientes; es decir, distribuciones con media cero y varianza uno. La forma de la distribución ji–cuadrado depende del número de sumandos libres involucrados en la suma. Por ejemplo, una ji–cuadrado conformada con k variables normales estándar independientes, escrita como

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2, \tag{2.16}$$

tiene una distribución ji–cuadrado que depende solo de k. En general, el número de sumandos libres asociados a una distribución ji–cuadrado se conoce como *grados de libertad* de la distribución. En la condición "tres números cuya suma sea 20", una vez que se han establecido dos sumandos (por ejemplo, 7 y 5), no hay libertad alguna para elegir el tercero (8); entonces, hay dos grados de libertad en dicha suma. Para el caso de tablas de contingencia de f–filas y c–columnas, el número de grados de libertad de la distribución es $(f-1)\times(c-1)$; por ejemplo, para la tabla 2.1 los grados de libertad son $(4-1)\times(3-1)=6$.

Para algunos grados de libertad y varios niveles de significancia se han elaborado tablas estadísticas donde se encuentran los respectivos valores de χ^2; de esta manera, se obtiene el valor de la respectiva distribución χ^2 (determinada por los grados de libertad) que deja un área a la derecha igual a la significancia asumida. La prueba estadística de la hipótesis de independencia se puede desarrollar también mediante la comparación de la estadística χ_0^2 (calculada desde los datos muestrales) con los valores tabulados para la distribución ji–cuadrado. De manera que si el valor χ_0^2, calculado mediante los datos, es mayor que el dispuesto en una tabla, para un número de grados de libertad y nivel de significancia establecidos, entonces se rechaza la hipótesis de independencia; en caso contrario, no se rechaza. Esta tabla corresponde a la (A.2) de los anexos.

En la actualidad, dado la disponibilidad de programas en estadística para computador (paquetes), el uso de las tablas ji–cuadrado ha sido desplazado, pues la decisión de rechazar o no la hipótesis de independencia se hace de acuerdo con el *valor p* suministrado por la salida de un paquete estadístico (figura 2.3); así, valores de p menores que 0.01 (1%) provocan el rechazo de la hipótesis nula, aunque a veces se admiten valores menores que 0.1 (10%) para tomar la misma decisión.

Para el caso que se ha venido tratando se desarrollan los cálculos conducentes a obtener el valor de la estadística χ^2, con el cual se verifica la hipótesis nula: "la opinión de las personas encuestadas es independiente de su estrato socieconómico".

Los valores para las frecuencias esperadas E_{ij} se calculan conforme a la fórmula (2.14); así, para la celda de la fila 1 y columna 1 ("bajo–bueno"), el valor de la frecuencia esperada E_{11} es dada por

$$\begin{aligned} E_{11} &= \frac{n_{1.}n_{.1}}{N} \\ &= \frac{150 \times 170}{500} \\ &= 51 \end{aligned}$$

Los demás valores E_{ij} (E_{12}, E_{13}, ..., E_{43}) se calculan en forma semejante. La tabla 2.7 contiene los resultados. El valor de la estadística χ^2

Tabla 2.7: Frecuencias esperadas.

	Opinión		
Estrato	Bueno	Regular	Malo
Bajo	(E_{11}) 51.0	(E_{12}) 48.0	(E_{13}) 51.0
Medio	(E_{21}) 61.2	(E_{22}) 57.6	(E_{23}) 61.2
Alto	(E_{31}) 30.6	(E_{32}) 28.8	(E_{33}) 30.6
Muy alto	(E_{41}) 27.2	(E_{42}) 25.6	(E_{43}) 27.2

se obtiene de calcular cada uno de los 12 valores de $(n_{ij} - E_{ij})^2/E_{ij}$ y luego sumarlos Estos cálculos se muestran en forma condensada a con-

tinuación:

$$\chi_0^2 = \frac{(75-51.0)^2}{51.0} + \frac{(35-48.0)^2}{48.0} + \cdots + \frac{(40-27.2)^2}{27.2}$$
$$\chi_0^2 = 11.294 + 3.520 + \cdots + 6.023$$
$$\chi_0^2 = 40.049$$

El valor de la variable χ^2, que se obtiene desde la tabla A.2 con 6 grados de libertad y un nivel de significación de $\alpha = 0.05$, es 12.592; en consecuencia, como el valor calculado para $\chi_0^2 = 40.049$ es mayor que el valor de la tabla, se rechaza la hipótesis de independencia entre las dos variables. Es decir, se puede afirmar, con un nivel de incertidumbre de 5%, que las variables no son independientes (están asociadas).

Un procedimiento alterno para apoyar la decisión estadística es la probabilidad de encontrar un valor mayor que o igual a $\chi_0^2 = 40.049$, en símbolos, $P(\chi^2 \geq 40.049) = p$. Este valor p, de acuerdo con el procedimiento FREQ del paquete SAS, es $p = 4.455 \times 10^{-7}$, de manera que el nivel de incertidumbre para rechazar la hipótesis de independencia es demasiado bajo, de donde se llega a la misma conclusión anterior.

En resumen, los datos evidencian una alta dependencia entre el estrato socioeconómico de los encuestados y su opinión sobre el sistema de salud.

2.4.3 Contraste mediante la razón de verosimilitudes (G^2)

La razón de verosimilitudes es el cociente de dos probabilidades: una es la probabilidad asociada a que el parámetro de una distribución sea el valor determinado por la hipótesis nula, otra es la probabilidad de que el parámetro proceda de un conjunto más general que el determinado por la hipótesis nula. De esta manera, la primera probabilidad es menor o igual que la segunda, pues el conjunto de valores del parámetro determinado por la hipótesis nula es un subconjunto del conjunto más general.

La estadística que se obtiene, después de un proceso de optimización, es:

$$G^2 = 2\sum_{i=1}^{f}\sum_{j=1}^{c} n_{ij} \ln\left(\frac{n_{ij}}{E_{ij}}\right) \tag{2.17}$$

Note que esta estadística es aritméticamente similar a la estadística χ^2, pues las diferencias $(n_{ij} - E_{ij})$ se corresponden con

$$\ln\left(\frac{n_{ij}}{E_{ij}}\right) = [\ln(n_{ij}) - \ln(E_{ij})]$$

Esta estadística se denomina *estadística de la razón de verosimilitudes ji–cuadrado*. Valores grandes de G^2 evidencian el rechazo que debe hacerse de la hipótesis nula (independencia). Para muestras grandes, la estadística G^2 tiene distribución ji–cuadrado con $(f-1)(c-1)$ grados de libertad. Algunos paquetes estadísticos como R o el SAS calculan esta estadística, para un conjunto de datos específico y reportan el valor p correspondiente.

Cuando se tiene la hipótesis de independencia, la estadística χ^2 y la estadística de razón de verosimilitud G^2 tienen distribución asintótica[1] ji–cuadrado con $(f-1)(c-1)$ grados de libertad. No es sencillo especificar el tamaño de muestra necesario para que la distribución ji–cuadrado aproxime adecuadamente las distribuciones exactas de χ^2 y G^2. Para un número fijo de celdas, χ^2 converge más rápidamente que G^2. La aproximación a la ji–cuadrado no es muy buena cuando $n/(f \times c) < 5$. Cuando el número de filas f o el número de columnas c es grande, puede ser adecuado la estadística χ^2 para $n/(f \times c)$ tan pequeños como 1, siempre que la tabla no contenga frecuencias esperadas altas ni bajas (Agresti 2000, 48–49).

Para los datos de la tabla 2.1, el valor de la estadística G^2 es

$$\begin{aligned} G^2 &= 2[n_{11}\ln(n_{11}/E_{11}) + n_{12}\ln(n_{12}/E_{12}) + \cdots + n_{43}\ln(n_{43}/E_{43})] \\ &= 2[40\ln(40/51.0) + 35\ln(35/48.0) + \cdots + 15\ln(15/27)] \\ &= 39.6925 \end{aligned}$$

Como $G^2 = 39.6925$, es mayor que $\chi^2_{(6,0.05)} = 12.59$ (tabla A.2), se concluye, nuevamente, que con estos datos las dos variables se muestran dependientes.

[1] Que convergen a la misma distribución, en este caso a la ji–cuadrado con $(f-1)(c-1)$ grados de libertad.

2.4.4 Medidas de asociación

Hasta ahora tan solo se ha analizado la existencia o no de una posible asociación entre dos variables de tipo categórico, pero no se ha establecido la *intensidad* de la posible relación. Una estadística que mida la magnitud de la asociación entre dos variables se denomina *medida de asociación* si la relación es *simétrica*, o sea que no hay distinción entre una variable de clasificación (variable explicativa o independiente) y una variable de respuesta (variable dependiente); se denomina *medida de tamaño del efecto* si la relación es asimétrica, es decir, una de las variables cumple el papel de variable respuesta.

A veces, se incurre en el error de emplear la estadística χ^2 como una medida totalmente concluyente sobre la relación, pues se puede llegar a interpretar como el grado en que los datos observados evidencian la existencia de una relación funcional entre las variables; así, valores grandes de χ_0^2 sugieren magnitudes grandes de asociación. No obstante, χ_0^2 puede aumentar su magnitud, artificialmente, conforme aumenta el tamaño de muestra, esto se evidencia a partir de la expresión (2.15a):

$$\chi_0^2 = N \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(p_{ij} - \widehat{\pi}_{i.}\widehat{\pi}_{.j})^2}{\widehat{\pi}_{i.}\widehat{\pi}_{.j}}$$

se observa que, por ejemplo, al duplicar el tamaño de muestra χ_0^2 también se duplica, pero la discrepancia entre los valores observados y esperados no se altera. Una medida de tamaño del efecto o magnitud de la asociación debe ser independiente del tamaño muestral, más cuando se pretende generalizar los resultados.

Se presentan aquí las medidas de uso frecuente que son reportadas en los paquetes R y SAS. Las más comunes son: el coeficiente de contingencia, la Q de Yule, el coeficiente *phi* (ϕ), la medida de asociación de Cramer (V), el coeficiente gama (γ), el coeficiente τ_b de Kendall, el coeficiente τ_c de Stuart, los coeficientes $D(X|Y)$ y $D(Y|X)$, el coeficiente de correlación de Pearson, el coeficiente de correlación de Spearman, el lambda asimétrico y el lambda simétrico. Es preciso advertir que estos son apenas indicadores (descriptivos), de manera que no deben ser considerados como soporte absoluto para garantizar o no la asociación entre variables. Se presentan a continuación, en forma condensada, estas estadísticas, las cuales son un apoyo para el esclarecimiento acerca el tipo y magnitud de una posible contingencia entre las variables de interés.

2.4.5 Medidas ligadas a la estadística ji–cuadrado

- *El coeficiente de contingencia.* Es una medida del grado de asociación o relación entre dos conjuntos de atributos. Es especialmente útil cuando se tiene información clasificadora acerca de uno o ambos conjuntos de atributos. El grado de asociación entre dos conjuntos de atributos, sean ordinales o no, se puede describir mediante la siguiente fórmula:

$$C = \sqrt{\frac{\chi_0^2}{\chi_0^2 + n}}, \text{ donde } \chi_0^2 = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (2.18)$$

  Note que la estadística C toma valores entre 0 y 1. Valores cercanos a cero muestran una baja asociación entre las variables, mientras que valores próximos a 1 dan cuenta de una posible alta asociación.

  El valor del coeficiente para el ejemplo considerado en este capítulo es

$$C = \sqrt{\frac{40.049}{40.049 + 500}} = 0.272$$

  Este coeficiente es invariante al orden de las modalidades de la variable fila o columna en una tabla de contingencia.

- *El coeficiente phi (ϕ).* Resulta de dividir el estadístico χ_0^2 por el tamaño muestral n y extraer su raíz cuadrada, es decir

$$\phi = \sqrt{\frac{\chi_0^2}{N}} = \sqrt{\sum_{i=1}^{f}\sum_{j=1}^{c} \frac{(p_{ij} - \widehat{\pi}_{i.}\widehat{\pi}_{.j})^2}{\widehat{\pi}_{i.}\widehat{\pi}_{.j}}} \qquad (2.19)$$

  Esta es una medida similar a la mostrada por la ecuación (2.18). Sin embargo, esta estadística puede no tener una cota superior cuando se calcula para tablas donde $f, c > 2$, en estos casos χ_0^2 puede ser mayor que el tamaño muestral n. Por esto, el uso de este coeficiente se restringe a tablas 2×2. No obstante, si $\phi = 0$ puede afirmarse con alta certidumbre que las variables son independientes, pero cuando $\phi > 0$ la interpretación no es concluyente.

  Para los datos de la tabla 2.1 el valor del coeficiente es

$$\phi = \sqrt{\frac{40.049}{500}} = 0.283$$

lo cual corrobora la dependencia entre estas variables.

- *El coeficiente (V) de Cramer.* Este coeficiente tiene un valor máximo en tablas de contingencia de cualquier tamaño, es igual al coeficiente phi cuando la tabla tiene un número de filas o de columnas igual a 2. Se define como

$$V = \sqrt{\frac{\chi_0^2}{nk}} \tag{2.20}$$

donde $k = \text{mín}\{f-1, c-1\}$ es el menor número de modalidades fila (o columna) menos uno de la tabla de contingencia.

El valor del coeficiente de Cramer para los datos de la tabla 2.1 es

$$V = \sqrt{\frac{\chi_0^2}{nk}} = \sqrt{\frac{40.049}{500(2)}} = 0.200$$

esto reafirma la conclusión de dependencia entre las variables.

2.4.6 Medidas basadas en la reducción proporcional del error (RPE)

Un inconveniente de las medidas de asociación que son función del estadístico χ^2 es que únicamente son válidas para estudios basados en el modelo de independencia (sobre todo en diseños transversales) con muestreo binomial o multinomial[2]. No obstante, la mayor dificultad estriba en la interpretabilidad: todas presentan el problema de que no parten de una conceptualización intuitiva de la correlación o asociación entre variables.

Una medida es la *reducción proporcional del error*, la cual se obtiene mediante la razón de una medida del error calculada como pronóstico de los valores de una variable y la misma medida de error obtenida como pronóstico de los valores de una variable adicional. Para ilustrar la lógica de la estadística, suponga que de una población se extraen personas al azar y se intenta "adivinar" qué modalidad les corresponde con relación a una variable de respuesta (Y). La predicción se establece de acuerdo con dos reglas: por la *primera*, no se emplea ninguna información previa

[2] Ato & López (1996, 51).

para pronosticar la respuesta; por la *segunda*, se examina a qué categoría pertenece el mismo individuo con relación a una variable de clasificación (X) y se determina o predice qué modalidad obtendrá en la variable Y.

Sea $P(1)$ la probabilidad de clasificar incorrectamente los individuos de acuerdo con la primera regla, y $P(2)$ la probabilidad de error bajo la segunda regla. La reducción de error que se obtiene con el empleo de la segunda regla, en oposición a la primera, es

$$RPE = \frac{P(1) - P(2)}{P(1)} \tag{2.21}$$

Estas medidas varían entre 0 y 1. Si las variables X y Y son independientes, entonces $P(1) = P(2)$, luego, $RPE = 0$. Por el contrario, si las variables X y Y tienen una asociación perfecta (dependientes), entonces $P(2) = 0$, y por tanto, $RPE = 1$; es decir, que el conocimiento de X permite pronosticar a la variable Y perfectamente.

A manera de ejemplo, considere los datos de la tabla 2.1. Sea Y la variable *opinión* sobre el servicio de salud y X la variable *estrato socioeconómico*. A continuación se transcriben los datos de la tabla 2.1 con las probabilidades entre paréntesis, las cuales resultan de dividir cada frecuencia (n_{ij}) entre el número total de individuos ($N = 500$). El es-

Tabla 2.8: Opinión sobre el servicio de salud.

	Opinión			
Estrato	Bueno	Regular	Malo	Total
Bajo	75	35	40	150
	(0.15)	(0.07)	(0.08)	(0.30)
Medio	60	70	50	180
	(0.12)	(0.14)	(0.10)	(0.36)
Alto	20	30	40	90
	(0.04)	(0.06)	(0.08)	(0.18)
Muy alto	15	25	40	80
	(0.03)	(0.05)	(0.08)	(0.16)
Total	170	160	170	500
	(0.34)	(0.32)	(0.34)	(1.00)

tadístico *lambda* (λ) se define en términos de las probabilidades $P(1)$ y $P(2)$ para cada una de las variables X y Y de la tabla de contingencia.

Asumiendo que la variable fila es la variable respuesta Y, se tiene que

$$P(1) = 1 - \sum_j \pi_{ij}^{\text{máx}}$$
$$P(2) = 1 - \pi_{i.}^{\text{máx}}$$

donde $P(1)$ suma los valores de probabilidad máximos por columna y $P(2)$ representa los marginales por fila. Si, por el contrario, la variable columna (opinión) se toma como variable respuesta Y, entonces

$$P(1) = 1 - \sum_i \pi_{ij}^{\text{máx}}$$
$$P(2) = 1 - \pi_{.j}^{\text{máx}}$$

donde $P(1)$ suma los valores de probabilidad máximos por fila y $P(2)$ representa los marginales por columna.

Los coeficientes resultantes de este intercambio de variables respuesta, $\lambda_{Y|X}$ y $\lambda_{X|Y}$ se estiman mediante

$$\widehat{\lambda}_{Y|X} = \frac{\sum_j p_{ij}^{\text{máx}} - p_{i.}^{\text{máx}}}{1 - p_{i.}^{\text{máx}}} \tag{2.22a}$$

$$\widehat{\lambda}_{X|Y} = \frac{\sum_i p_{ij}^{\text{máx}} - p_{.j}^{\text{máx}}}{1 - p_{.j}^{\text{máx}}} \tag{2.22b}$$

Note que $p_{i.}^{\text{máx}}$ y $p_{.j}^{\text{máx}}$ son las mayores probabilidades marginales de las filas y las columnas, respectivamente.

A estos coeficientes también se les denomina *lambda asimétricos*. Así, $\lambda_{Y|X}$ se interpreta como el probable incremento de predecir la variable columna Y dado que hay conocimiento sobre la variable fila X; una interpretación similar tiene $\lambda_{X|Y}$. Estos coeficientes tienen alguna semejanza con el *coeficiente de determinación* para el caso de la regresión con variables en escala o de intervalo.

La cantidad $\sum_j p_{ij}^{\text{máx}}$ corresponde a la suma de las probabilidades más grandes en cada columna; de manera semejante, $\sum_i p_{ij}^{\text{máx}}$ es la suma de las probabilidades más grandes en cada fila.

Para los datos de la tabla 2.1, se tienen las siguientes estimaciones de

los lambda:

$$\lambda_{Y|X} = \frac{(0.15 + 0.14 + 0.10) - 0.36}{1 - 0.36} = \frac{0.03}{0.64} = 0.047$$
$$\lambda_{X|Y} = \frac{(0.15 + 0.14 + 0.08 + 0.08) - 0.34}{1 - 0.34} = \frac{0.11}{0.66} = 0.167$$

Para el caso $\lambda_{Y|X}$, el valor 0.15 es la máxima probabilidad en la primera columna; 0.14 en la segunda; 0.10 en la tercera (última columna); y 0.36 es la probabilidad marginal más grande por filas. De manera similar, para el cálculo de $\lambda_{X|Y}$, 0.15 es la probabilidad más grande en la primera fila; 0.14 en la segunda; 0.08 en la tercera, 0.08 en la cuarta (última fila); y 0.34 es la probabilidad marginal más grande entre las columnas.

Los estadísticos anteriores se denominan asimétricos porque al intercambiar la variable respuesta con la explicativa, el valor de los estadísticos no siempre es igual. En modelos de independencia no hay distinción entre variable respuesta e independiente, luego se requiere una versión *simétrica* del coeficiente lambda. Esta se define por

$$\widehat{\lambda} = \frac{\sum_i p_{ij}^{\text{máx}} + \sum_j p_{ij}^{\text{máx}} - p_{i.}^{\text{máx}} - p_{.j}^{\text{máx}}}{2 - p_{i.}^{\text{máx}} - p_{.j}^{\text{máx}}} \tag{2.23}$$

Este coeficiente toma valores entre 0.0 y 1.0. Valores cercanos a 1.0 evidencian una alta asociación entre las variables.

Para los datos de la tabla 2.8, el valor de la estadística $\widehat{\lambda}$ es

$$\widehat{\lambda} = \frac{(0.15 + 0.14 + 0.08 + 0.08) + (0.15 + 0.14 + 0.10) - 0.36 - 0.34}{2 - 0.36 - 0.34}$$
$$= \frac{0.14}{1.3} = 0.1076$$

lo cual advierte sobre la existencia de una asociación baja entre estas variables.

2.4.7 Medidas de asociación ordinales

Variables cuyas modalidades tienen un orden determinado requieren estadísticas que incorporen o consideren esta información al análisis de asociación. Para el caso considerado, las variables estrato socioeconómico y opinión tienen modalidades que siguen un orden.

En este tipo de variables se puede buscar una asociación con tendencia. Por ejemplo, se quiere establecer si, cuando aumenta el estrato, la proporción de personas ubicadas en los niveles de opinión alto también aumenta o, por el contrario, las respuestas en los niveles altos tienden a decrecer.

Se ha propuesto una serie de estadísticas con las cuales se puede contrastar la hipótesis nula de "independencia" frente a la hipótesis alternativa de "tendencia lineal". La estadística de aplicación más frecuente, con la cual se verifica la hipótesis alternativa de que existe una asociación lineal entre la variable fila y la variable columna, es la ji–cuadrado de Manthel–Haenszel (Agresti 1996, 34–39), dada por la expresión

$$M^2 = (n-1)r^2 \tag{2.24}$$

donde r^2 es el coeficiente de correlación de Pearson entre las variables fila y columna. Como en el caso de la prueba ji–cuadrado, la decisión de rechazar la hipótesis de independencia frente a la alternativa de tendencia lineal se guía por el valor p de M^2, esta estadística tiene una distribución ji–cuadrado con un grado de libertad.

Cuando se rechaza la hipótesis de no correlación se puede optar por la hipótesis de tendencia lineal. En este punto el problema ahora es qué tipo de tendencia lineal evidencian los datos, porque esta tendencia puede ser creciente o decreciente. Algunas estadísticas útiles para apoyar la afirmación sobre el tipo de tendencia se esquematizan a continuación.

- *Estadística gama ($\widehat{\gamma}$)*

  Cuando se considera el orden de un individuo respecto a dos variables ordinales, A y B, es posible determinar por comparación si el par es *concordante*, es decir, cuando un sujeto con puntuación alta en A también tiene puntuación alta en B (y recíprocamente), o *discordante* cuando un sujeto que puntúa alto en A puntúa bajo en B. El par se dice *empatado* si el sujeto obtiene la misma puntuación en A y B. La estadística que da cuenta del tipo de asociación lineal entre dos variables es la *gama*, $\widehat{\gamma}$, está definida por

  $$\widehat{\gamma} = \frac{(C-D)}{(C+D)} \tag{2.25}$$

  donde C es el número de pares concordantes y D el número de pares discordantes en la muestra de individuos. Esta estadística

toma valores entre -1 y $+1$. Valores de $\widehat{\gamma}$ cercanos a $+1$ sugieren una tendencia lineal creciente (positiva) entre las dos variables, mientras que valores de $\widehat{\gamma}$ cercanos a -1 advierten sobre una tendencia lineal decreciente (negativa); valores próximos a cero no permiten decidir sobre asociación lineal significativa. De la definición de este coeficiente se observa que es simétrico, de manera que un cambio en la ordenación de las variables cambia el signo de $\widehat{\gamma}$.

Cuando se dispone únicamente de los datos en una tabla de contingencia, el cálculo de los valores de C y D puede establecerse definiendo primero la frecuencia total del segundo elemento del par y multiplicando después por la frecuencia del primer elemento del par. De esta manera, es conveniente calcular previamente la frecuencia total del segundo elemento del par mediante

$$a_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} \tag{2.26a}$$

$$b_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} \tag{2.26b}$$

donde a_{ij} y b_{ij} son las frecuencias totales de casos concordantes y discordantes del segundo elemento del par que corresponden a las celdas (i, j). El valor de C y D se obtiene a través de

$$C = \sum_{i} \sum_{j} n_{ij} a_{ij} \tag{2.27a}$$

$$D = \sum_{i} \sum_{j} n_{ij} b_{ij} \tag{2.27b}$$

En la tabla 2.1, en la variable columna (opinión) las modalidades no están ordenadas de manera creciente como la variable fila; una readecuación de esta tabla, con el ánimo de ilustrar los cálculos, se muestra en la tabla 2.9. A continuación se describen los cálculos que conducen al valor de la estadística gama. Los valores de las concordancias C y las discordancias D, de acuerdo con las expresiones (2.27a) y (2.27b) se muestran en las tablas 2.10 y 2.11, respectivamente. El valor de la estadística gama, de acuerdo con

Tabla 2.9: Opinión sobre el servicio de salud (de tabla 2.1).

	Opinión			
Estrato	Malo	Regular	Bueno	Total
Bajo	40	35	75	150
Medio	50	70	60	180
Alto	40	30	20	90
Muy alto	40	25	15	80
Total	170	160	170	500

Tabla 2.10: Concordancias.

Celda (i, j)	$n_{ij}[\sum_{k>i}\sum_{l>j} n_{kl}]$	Total
1-1	40[70+60+30+20+25+15]	8800
1-2	35[60+20+15]	3325
1-3	75[0]	0
2-1	50[30+20+25+15]	4500
2-2	70[20+15]	2450
2-3	60[0]	0
3-1	40[25+15]	1600
3-2	30[15]	450
3-3	20[0]	0
4-1	40[0]	0
4-2	25[0]	0
4-3	15[0]	0
		C=21125

la expresión (2.25), es

$$\widehat{\gamma} = \frac{(C-D)}{(C+D)} = \frac{21125 - 39875}{21125 + 39875}$$
$$= \frac{-18750}{61000} = -0.3073$$

De esta manera se podría deducir que se evidencia una relación inversa entre el nivel socioeconómico y la opinión favorable sobre el servicio de salud. Esta conclusión se sustentará y ampliará cuando se determinen las fuentes de asociación (sección 2.4.9).

Tabla 2.11: Discordancias.

Celda (i,j)	$n_{ij}[\sum_{k>i}\sum_{l<j} n_{kl}]$	Total
1-1	40[0]	0
1-2	35[50+40+40]	4550
1-3	75[50+70+40+30+40+25]	19125
2-1	50[0]	0
2-2	70[40+40]	5600
2-3	60[40+30+40+25]	8100
3-1	40[0]	0
3-2	30[40]	1200
3-3	20[40+25]	1300
4-1	40[0]	0
4-2	25[0]	0
4-3	15[0]	0
		D=39875

- *Estadística Tau_b (τ_b) de Kendall.*

  Es similar al coeficiente gama, excepto que este emplea una corrección para los empates (Kendall & Stuart 1979). Es apropiado cuando las dos variables están en una escala ordinal, toma valores que varían entre -1 y $+1$, la interpretación es semejante a la del coeficiente gama. Se define como

$$\widehat{\tau}_b = \frac{C - D}{\sqrt{[W_f W_c]}} \tag{2.28}$$

  donde W_f y W_c se calculan mediante

$$W_f = \frac{N^2 - \sum_i n_{i.}^2}{2} \tag{2.29a}$$

$$W_c = \frac{N^2 - \sum_j n_{.j}^2}{2} \tag{2.29b}$$

Para los datos de la tabla 2.1, el valor de la estadística $\widehat{\tau}_b$ se calcula

conforme a las expresiones (2.28) y (2.29a-b), así:

$$W_f = \frac{N^2 - \sum_i n_{i.}^2}{2} = \frac{[500^2 - (150^2 + 180^2 + 90^2 + 80^2)]}{2}$$
$$= 90300$$
$$W_c = \frac{N^2 - \sum_j n_{.j}^2}{2} = \frac{[500^2 - (170^2 + 160^2 + 170^2)]}{2}$$
$$= 83300$$

luego

$$\widehat{\tau}_b = \frac{C - D}{\sqrt{[W_f W_c]}} = \frac{-18750}{\sqrt{(90300)(83300)}} = -0.2162$$

- ! *Estadística Tau_c ($\widehat{\tau}_c$) de Stuart.*

  Otra medida de asociación afín con la anterior es el coeficiente Tau_c, el cual se define mediante

$$\widehat{\tau}_c = \frac{C - D}{N^2\left(\frac{k-1}{2k}\right)} \tag{2.30}$$

  con k el número menor de filas o columnas de la tabla de contingencia, $k = \text{mín}\{f, c\}$. Para los datos sobre la opinión del servicio de salud el coeficiente Tau_c toma el valor:

$$\widehat{\tau}_c = \frac{C - D}{N^2\left(\frac{k-1}{2k}\right)} = \frac{-18750}{500^2\left(\frac{3-1}{2(3)}\right)} = -0.225$$

- *Estadística de Somers ($D(Y|X)$).*

  Es una "medida" asimétrica, la cual es una modificación del $\widehat{\tau}_b$. Como en la estadística lambda, $Y|X$ denota que la variable X se considera variable estadísticamente explicativa (variable independiente) de la variable Y (variable dependiente), similarmente, $X|Y$ denota que Y es la variable explicativa de la variable X.

  Una diferencia del coeficiente de Somers con el coeficiente $\widehat{\tau}_b$ es que usa solo corrección para los empates en la variable explicativa. El coeficiente D de Somers es apropiado únicamente cuando las variables están, en al menos, escala ordinal. Su rango está entre

-1 y $+1$. La fórmula del coeficiente de Somers es

$$D(Y|X) = \frac{C - D}{W_f}$$
$$D(X|Y) = \frac{C - D}{W_c}$$

con C, D, W_f y W_c definidos en (2.27a–b) y (2.29a–b), respectivamente. Los valores de la estadística $D(|)$ para los datos de la tabla 2.8 son

$$D(Y|X) = \frac{C - D}{W_f} = \frac{-18750}{90300} = -0.2076$$
$$D(X|Y) = \frac{C - D}{W_c} = \frac{-18750}{83300} = -0.2250$$

El valor de $D(Y|X) = -0.2076$ muestra que existe alguna asociación (inversa) entre la condición socioeconómica y la apreciación de los encuestados respecto al servicio de salud. La interpretación de $D(X|Y) = -0.2250$ es semejante, aunque es el investigador quien establece la variable que cumple el papel de explicativa y la que se considera dependiente.

2.4.8 Otras medidas de asociación

Uno de los coeficientes de asociación más importantes, del cual por generalización se pueden derivar otros análisis como el canónico y el de correspondencias (Díaz 2007), es el *coeficiente de correlación de Pearson (ρ_p)*.

- *Coeficiente de correlación de Pearson (ρ_p).*

  El coeficiente de correlación de Pearson se computa mediante puntajes asignados a cada una de las modalidades de las dos variables. El rango de ρ_p está entre -1 y $+1$. El coeficiente se calcula mediante

$$\widehat{\rho}_p = \frac{\sum_i \sum_j n_{ij}(f_i - \bar{f})(c_j - \bar{c})}{\sqrt{\left[\sum_i \sum_j n_{ij}(f_i - \bar{f})^2\right]\left[\sum_i \sum_j n_{ij}(c_j - \bar{c})\right]^2}} \qquad (2.31)$$

donde f_i y c_j son los puntajes para las modalidades de las variables fila y columna, respectivamente con

$$\bar{f} = \frac{\sum_i n_{i.} f_i}{n} \quad \text{y} \quad \bar{c} = \frac{\sum_j n_{.j} c_j}{n} \tag{2.32}$$

Para los datos de la tabla 2.8, los puntajes para la variable estrato (fila) son 1, 2, 3 y 4, mientras que para la variable opinión (columna) son 1, 2 y 3; aunque pueden ser otros valores. De acuerdo con la ecuación (2.32), $\bar{f} = 2.2$ y $\bar{c} = 2.0$. El coeficiente de asociación de Pearson para medir la correlación entre la opinión sobre el servicio de salud y el estrato socieconómico de las 500 personas encuestadas es

$$\begin{aligned}\widehat{\rho}_p &= \frac{\sum_i \sum_j n_{ij}(f_i - \bar{f})(c_j - \bar{c})}{\sqrt{\left[\sum_i \sum_j n_{ij}(f_i - \bar{f})^2\right]\left[\sum_i \sum_j n_{ij}(c_j - \bar{c})\right]^2}} \\ &= \frac{-105}{\sqrt{(540)(340)}} = -0.2450\end{aligned}$$

con el cual se derivan conclusiones semejantes a las anteriores.

- *Coeficiente de correlación de Spearman (ρ_s).*

  Es apropiado cuando las variables están solo en escala ordinal. Este coeficiente se calcula mediante la misma fórmula del coeficiente de Pearson, empleando los puntajes f_i y c_j, calculados de acuerdo con las siguientes expresiones:

$$\begin{aligned}f_i &= \sum_{k<i} n_{k.} + \frac{(n_{i.}+1)}{2} \; i = 1, 2, \ldots, f \text{ y} \\ c_j &= \sum_{l<j} n_{.l} + \frac{(n_{.j}+1)}{2} \; j = 1, 2, \ldots, c\end{aligned} \tag{2.33}$$

  Para los datos considerados, los puntajes fila son $f_1 = 75.5$, $f_2 = 240.5$, $f_3 = 375.5$ y $f_4 = 460.5$ y los puntajes columna son: $c_1 = 85.5$, $c_2 = 250.5$ y $c_3 = 415.5$. El valor del coeficiente de Spearman es

$$\begin{aligned}\widehat{\rho}_s &= \frac{\sum_i \sum_j n_{ij}(f_i - \bar{f})(c_j - \bar{c})}{\sqrt{\left[\sum_i \sum_j n_{ij}(f_i - \bar{f})^2\right]\left[\sum_i \sum_j n_{ij}(c_j - \bar{c})\right]^2}} \\ &= \frac{-2305875}{\sqrt{(9546000)(9256500)}} = -0.2453\end{aligned}$$

- *Coeficiente de incertidumbre ($U(Y|X)$).*

  Es una medida análoga a las medidas de varianza no explicada utilizada en modelos de regresión o en diseños experimentales (entropía). El coeficiente de *incertidumbre* registra la proporción en que se reduce la incertidumbre de una variable respuesta como resultado del conocimiento del valor de otra variable. Este es un coeficiente asimétrico, cuyo rango está entre 0.0 y 1.0, valores cercanos a 1.0 dan cuenta de una alta asociación o covariación entre las variables, mientras que valores cercanos a 0 sugieren que las variables son independientes. Un problema de esta medida, que también lo tienen las anteriores, es la carencia de una medida que permita afirmar qué valores definen una asociación relevante entre las variables. El coeficiente de incertidumbre de la variable Y con relación a la variable X se define como

$$U(Y|X) = \frac{[H(X) + H(Y) - H(XY)]}{H(Y)} \tag{2.34}$$

  donde

$$\begin{aligned} H(X) &= -\sum_i (n_{i.}/n) \ln(n_{i.}/n) \\ H(Y) &= -\sum_j (n_{.j}/n) \ln(n_{.j}/n) \\ H(XY) &= -\sum_i \sum_j (n_{ij}/n) \ln(n_{ij}/n) \end{aligned} \tag{2.35}$$

  El coeficiente $U(X|Y)$ se calcula de manera similar por intercambio de los índices. Para los datos anteriores, si se considera variable respuesta la opinión y variable explicativa el estrato, entonces el valor de este coeficiente se calcula como sigue:

$$\begin{aligned} H(X) = -\Bigg[&\frac{150}{500}\ln\left(\frac{150}{500}\right) + \frac{180}{500}\ln\left(\frac{180}{500}\right) + \frac{90}{500}\ln\left(\frac{90}{500}\right) \\ &+\frac{80}{500}\ln\left(\frac{80}{500}\right)\Bigg] = 1.3309 \end{aligned}$$

$$\begin{aligned} H(Y) = -\Bigg[&\frac{170}{500}\ln\left(\frac{170}{500}\right) + \frac{160}{500}\ln\left(\frac{160}{500}\right) \\ &+\frac{170}{500}\ln\left(\frac{170}{500}\right)\Bigg] = 1.0982 \end{aligned}$$

$$H(XY) = -\left[\frac{40}{500}\ln\left(\frac{40}{500}\right) + \frac{35}{500}\ln\left(\frac{35}{500}\right) + \cdots + \frac{15}{500}\ln\left(\frac{15}{500}\right)\right] = 2.3893$$

así,

$$U(Y|X) = \frac{[H(X) + H(Y) - H(XY)]}{H(Y)} = \frac{[1.3309 + 1.0982 - 2.3893]}{1.0982} = 0.0361.$$

Mediante cálculos similares $U(X|Y) = 0.0298$; note que estos valores son bajos. Se puede afirmar entonces que la información de la variable respuesta que no es "explicada" por la variable independiente es, relativamente, baja.

Una versión *simétrica* de este coeficiente es

$$U = 2\frac{[H(X) + H(Y) - H(XY)]}{[H(X) + H(Y)]} \tag{2.36}$$

Para los datos del ejemplo, $U = 0.0327$. Al final de este capítulo se muestra la sintaxis del paquete SAS del procedimiento FREQ, con la cual se obtienen las estadísticas presentadas hasta ahora.

2.4.9 Determinación de las fuentes de asociación

En el análisis de las tablas anteriores, las pruebas proporcionan la evidencia de la existencia o no de la independencia entre las variables que se estudian. No obstante, el investigador no dispone de información para determinar en mayor medida cuáles son las modalidades responsables de la no independencia. Este es un problema similar al que se presenta en el análisis de varianza para la comparación de medias, ya que cuando la prueba detecta que hay diferencias entre las medias, se efectúan comparaciones entre algunas medias para esclarecer la(s) fuente(s) de la diferencia.

Para la detección de las fuentes (modalidades) de asociación se dispone de dos estrategias: una denominada *directa* y otra *de conversión a tablas bidimensionales*. En el primer caso está el *análisis de residuales*; en el segundo, la partición de la tabla original en tablas 2×2.

2.4.10 Análisis de los residuos

Es un procedimiento semejante al empleado en análisis de regresión, el cual consiste en el estudio de los residuales, es decir, el análisis de las diferencias entre los valores empíricos observados (n_{ij}) y los valores estimados (E_{ij}). Cada una de las diferencias $e_{ij} = n_{ij} - E_{ij}$ se denomina residual. Desde el punto de vista del análisis, es más conveniente el estudio de los residuales *tipificados*, definidos por

$$e_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}} \tag{2.37}$$

Note que la suma de los cuadrados de los e_{ij} es igual al valor de la estadística χ^2. Con esta medida se puede considerar también la contribución de cada celda al estadístico χ^2.

Los residuales definidos en (2.37) tienen varianza casi siempre menor que la unidad; más bien tienen varianza asintóticamente igual a uno. Un análisis más adecuado se logra con los residuales *ajustados*, que corresponden a la razón entre el residual tipificado y su error típico estimado, es decir:

$$d_{ij} = \frac{(n_{ij} - E_{ij})/\sqrt{E_{ij}}}{\sqrt{v_{ij}}} \tag{2.38}$$

donde v_{ij} es la varianza de los e_{ij}. Esta se define por

$$v_{ij} = \text{var}(e_{ij}) = \Big(1 - \frac{n_{i.}}{N}\Big)\Big(1 - \frac{n_{.j}}{N}\Big) \tag{2.39}$$

Cuando las dos variables que conforman la tabla de contingencia son independientes, cada residual ajustado d_{ij} se distribuye aproximadamente como una normal estándar (una $n(0,1)$). En consecuencia, mediante esta propiedad se pueden buscar las celdas que provocan el rechazo de la hipótesis, comparando los valores absolutos de los residuales ajustados con el correspondiente cuantil de la normal estándar, para un nivel de significancia específico. Las celdas que muestran una falta de ajuste entre los valores observados y los valores esperados y que, por tanto, hacen que se rechace la hipótesis nula de independencia, advierten sobre las modalidades de las respectivas variables a las que se les puede atribuir la asociación evidenciada.

En el análisis de los datos del nivel socioeconómico con el grado de satisfacción sobre el servicio de salud se verificó que estas variables se asocian. No obstante, se quiere conocer las fuentes de dicha dependencia.

Para esto se calculan los residuales tipificados (con los datos de las tablas (2.1) y (2.7)) e_{ij} y sus varianzas v_{ij}; así, para la celda $1-1$:

$$\begin{aligned} e_{11} &= \frac{n_{11} - E_{11}}{\sqrt{E_{11}}} \\ &= \frac{75 - 51.0}{\sqrt{51.0}} \\ &= 3.3607 \end{aligned}$$

La varianza de este residual, de acuerdo con (2.39), es

$$\begin{aligned} v_{11} &= \Big(1 - \frac{n_{1.}}{N}\Big)\Big(1 - \frac{n_{.1}}{N}\Big) \\ &= \Big(1 - \frac{150}{500}\Big)\Big(1 - \frac{170}{500}\Big) \\ &= 0.4620 \end{aligned}$$

el residual d_{11} ajustado es igual a

$$\begin{aligned} d_{11} &= \frac{e_{11}}{\sqrt{v_{11}}} \\ &= \frac{3.3607}{\sqrt{0.4620}} \\ &= 4.9444 \end{aligned}$$

Los demás valores se calculan de manera semejante. Estos se consignan en la tabla 2.12. Para un nivel de significancia de 5% ($Z_{0.025} = 1.96$) se observa, en la tabla 2.12, que varios de los residuales ajustados son significativos (mayores que 1.96). De estos residuales se observa un número considerablemente mayor de encuestados que pertenecen a una clase baja y poseen una opinión favorable sobre el servicio de salud que reciben. De esta manera se puede deducir no solo que existe una relación, sino, además, que esta va en dirección contraria al estrato socieconómico.

Esta metodología supone un análisis para cada una de las celdas. A diferencia del contraste χ^2 que supone $(f-1)\times(c-1)$ elementos independientes, el análisis de residuos, d_{ij}, implica que la totalidad de estos $(f \times c)$ son independientes, y que cada uno de ellos se ajusta a una distribución normal estándar $(n(0,1))$.

Tabla 2.12: Residuales.

e_{ij}		v_{ij}		d_{ij}	
3.3607	1.6930	0.4620	0.5412	4.9444	2.3098
-0.1534	2.4542	0.4224	0.5544	-0.2360	3.2960
-1.9162	-1.8764	0.5412	0.7176	-2.6046	-2.7198
-2.3392	1.6338	0.5544	0.4352	-3.1416	2.4766
-1.5403	0.2236	0.4620	0.5576	-2.2661	0.2995
-1.4316	-0.1186	0.4224	0.5712	-2.2028	-0.1569

2.4.11 Partición de tablas

Es otro procedimiento para detectar fuentes de asociación. La metodología consiste en simplificar la tabla original en varias tablas bidimensionales (2×2). La partición se sustenta en la propiedad de la distribución χ^2, según la cual *la suma de k variables aleatorias independientes con distribución* χ^2 *de* v_i *grados de libertad, es también* χ^2 *con* $v = v_1 + v_2 + \cdots + v_k$ *grados de libertad*. La aplicación de esta propiedad permite descomponer una tabla de contingencia, de tamaño $f \times c$, en varias subtablas, cada una de tamaño 2×2 y calcular una prueba χ^2 para cada una de ellas. El empleo de la estadística χ^2 plantea el inconveniente que la suma de variables aleatorias independientes no sea exactamente igual al valor del estadístico. Sin embargo, la utilización de la estadística de la razón de verosimilitudes (G^2) permite superarlo (Ato & López 1996, 42–44).

Para la formación de subtablas se han sugerido unas reglas con las cuales se consiguen subtablas que producen componentes independientes de G^2. Estas son:

1. La suma de los grados de libertad de las diferentes subtablas debe ser igual a los grados de libertad de la tabla original.

2. Las frecuencias de cada celda de la tabla original deben aparecer en una y solo una de las diferentes subtablas.

3. Cada total marginal de la tabla original debe aparecer una sola vez en cada una de las subtablas.

Un esquema de partición adecuado para formar tablas 2×2 es el siguiente:

$\sum_{a<i} \sum_{b<j} n_{ab}$	$\sum_{a<i} n_{aj}$
$\sum_{b<j} n_{ib}$	n_{ij}

Para los datos de la encuesta sobre la opinión del servicio de salud, de acuerdo con el esquema anterior, se deben conformar las siguientes tablas 2×2.

Subtabla	Frecuencias		Total	G^2	(%)
1.	$n_{11} = 40$	$n_{12} = 35$	75		
	$n_{21} = 50$	$n_{22} = 70$	120		
Total	90	105	195	2.528	(6.3)
2.	$n_{11} + n_{12} = 75$	$n_{13} = 75$	150		
	$n_{21} + n_{22} = 120$	$n_{23} = 60$	180		
Total	195	135	330	9.418*	(23.7)
3.	$n_{11} + n_{21} = 90$	$n_{12} + n_{22} = 105$	195		
	$n_{31} = 40$	$n_{32} = 30$	70		
Total	130	135	265	2.494	(6.2)
4.	$n_{11} + n_{12} + n_{21} + n_{22} = 195$	$n_{13} + n_{23} = 135$	330		
	$n_{31} + n_{32} = 70$	$n_{33} = 20$	90		
Total	265	155	420	11.241*	(28.3)
5.	$n_{11} + n_{21} + n_{31} = 130$	$n_{12} + n_{22} + n_{32} = 135$	265		
	$n_{41} = 40$	$n_{42} = 25$	65		
Total	170	160	330	3.284	(8.2)
6.	$n_{11} + n_{12} + n_{21} + n_{22} + n_{31} + n_{32} = 265$	$n_{13} + n_{23} + n_{33} = 155$	420		
	$n_{41} + n_{42} = 65$	$n_{43} = 15$	80		
Total	330	170	500	10.728*	(27.0)
			TOTAL	39.693	(100.0)

**: Se interpreta como asociación significativa al 5%.*

La proporción de G^2 más importante (28.3% del valor total de G^2) corresponde a la subtabla 4, la cual se obtiene de combinar las frecuencias de los dos estratos bajos con las respuestas desfavorables y regulares frente a las respuestas de favorabilidad en los estratos altos. Le sigue en importancia (con un 27.0% de G^2) la subtabla 6 que colapsa los tres primeros estratos con sus respuestas de malo y regular para contraponerlas con las respuestas de favorabilidad del estrato socieconómicamente más alto. Un resultado similar se observa en la subtabla 2.. En los tres casos se nota la asociación entre las respuestas de favorabilidad, positiva y negativa, determinada por los estratos bajo y alto, respectivamente.

Este resultado es semejante al encontrado con los procedimientos presentados.

Una limitación de este procedimiento es que no es apropiado para responder a todas las posibles asociaciones entre dos variables.

2.4.12 Análisis con el PROC FREQ del paquete estadístico SAS

El procedimiento FREQ produce tablas de una vía de clasificación a tablas a n vías de clasificación cruzada. Para tablas de doble vía, el procedimiento computa pruebas y medidas de asociación. Para tablas a n vías de clasificación con la rutina FREQ se puede desarrollar un análisis por estratos; mediante el cual se calculan algunas estadísticas dentro de los estratos y entre estos.

A continuación se muestra el esquema general de las instrucciones requeridas por el procedimiento FREQ. Las palabras de uso obligatorio para los comandos se escriben en mayúsculas fijas (esto nada más que por simple presentación); los comentarios sobre cada instrucción se escriben dentro de los símbolos /∗ y ∗/.

```
PROC FREQ opciones;
TABLES variables / opciones;
/* las tablas de dos o más entradas se escriben conectando
 la variables con el símbolo * */
WEIGHT variable; /* aquí se usa para entrar la frecuencia
                    de celda */
BY variable;
/* para obtener análisis separado en las
   variables indicadas en el BY */
RUN; /* para ejecutar el procedimiento */
```

A continuación se presenta la sintaxis y la salida del procedimiento FREQ del SAS, con el cual se desarrollan los cálculos involucrados en los análisis hasta aquí tratados, para los datos de la tabla de contingencia (2.1).

```
DATA tabla2_1;
/*Se nombra el archivo como tabla2_1.
```

```
El ';' es obligatorio al final de cada instrucción*/
INPUT estrato$ opinion$ frecue;
/*Declara las variables fila y columna y la frecuencia
  de celdas*/
/*El $ Señala una variable categórica. frecue
  corresponde  a la frecuencia*/
CARDS; /*Anuncia  los datos*/
/*Se ingresan los datos conforme al orden de las
  variables del INPUT */
est1 malo 40 est1 regular 35 est1 super 75 est2 malo 50
est2 regular 70 est2 super 60 est3 malo 40 est3 regular
30 est3 super 20 est4 malo 40 est4 regular 25 est4
super 15
;
PROC FREQ; /*Invoca el procedimiento FREQ*/
WEIGHT frecue; /* Indica que la frecuencia de cada
celda es el valor de FRECUE */
TABLES   estrato*opinion / CHISQ MEASURES DEVIATION;
             /*Filas y columnas de la tabla*/
/*/CHISQ pide la estadística ji--cuadrado, la G^2,
 Mantel-Haenszel, phi,  de contingencia y de Cramer*/
/*La opción MEASURE produce las demás medidas de
  asociación y su error estándar asintótico (ASE)*/
/*La opción DEVIATION produce los residuales de la
  forma n_{ij}-E_{ij}  */
RUN;
```

La tabla 2.13 reúne cinco tablas en una sola, es decir, las tablas con las frecuencias absolutas, los residuales, los porcentajes (frecuencias relativas) y los perfiles fila y columna, respectivamente.

2.5 Tablas de contingencia 2×2

Las tablas de contingencia 2×2, no obstante ser las más sencillas, son las de uso más frecuente en investigación médica, social y educación.

Un análisis sobre este tipo de tablas corresponde a la exploración y búsqueda de la posible asociación entre dos variables categóricas dicotómicas (o dicotomizadas). Por simplicidad, se notan las entradas de

Tabla 2.13: Salida SAS.

ESTRATO / Frequency / Deviation / Percent / Row Pct / Col Pct	OPINION			
	malo	regular	super	Total
est1	40	35	75	150
	-11	-13	25	
	8.00	7.00	15.00	30.00
	26.67	23.33	50.00	
	23.53	21.88	44.12	
est2	50	70	60	180
	-11.2	12.4	-1.2	
	10.00	14.00	12.00	36.00
	27.78	38.89	33.33	
	29.41	43.75	35.29	
est3	40	30	20	90
	9.4	1.2	-10.6	
	8.00	6.00	4.00	18.00
	44.44	33.33	22.22	
	23.53	18.75	11.76	
est4	40	25	15	80
	12.8	-0.6	-12.2	
	8.00	5.00	3.00	16.00
	50.00	31.25	18.75	
	23.53	15.63	8.82	
Total	170	160	170	500
	34.00	32.00	34.00	100.00

STATISTICS FOR TABLE OF ESTRATO BY OPINION

Statistic	DF	Value	Prob
Chi-Square	6	40.049	0.001
Likelihood Ratio Chi-Square	6	39.693	0.001
Mantel-Haenszel Chi-Square	1	29.964	0.001
Phi Coefficient		0.283	
Contingency Coefficient		0.272	
Cramer's V		0.200	

estas tablas con las primeras cuatro letras del alfabeto, así: $a = n_{11}$, $b = n_{12}$, $c = n_{21}$ y $d = n_{22}$. La tabla 2.14 muestra las frecuencias asociadas a estas tablas. Se presenta la estadística ji–cuadrado y algunas

Statistic	Value	ASE
Gamma	-0.307	0.054
Kendall's Tau-b	-0.216	0.039
Stuart's Tau-c	-0.225	0.040
Somers' D $C\|R$	-0.208	0.037
Somers' D $R\|C$	-0.225	0.040
Pearson Correlation	-0.245	0.043
Spearman Correlation	-0.245	0.044
Lambda Asymmetric $C\|R$	0.167	0.042
Lambda Asymmetric $R\|C$	0.047	0.035
Lambda Symmetric	0.108	0.033
Uncertainty Coefficient $C\|R$	0.036	0.011
Uncertainty Coefficient $R\|C$	0.030	0.009
Uncertainty Coefficient Symmetric	0.033	0.010

Sample Size = 500

estadísticas para tablas especiales en el análisis de asociación entre la variable fila y columna de tales tablas.

Tabla 2.14: Tabla de contingencia 2×2.

	Variable columna		
Variable fila	Categoría 1	Categoría 2	Total
Categoría 1	a	b	a+b
Categoría 2	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

2.5.1 Prueba ji–cuadrado

La estadística ji–cuadrado puede expresarse en términos de las frecuencias mostradas en la tabla 2.14 y desde la fórmula (2.15b). Esta es:

$$\chi_0^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad (2.40)$$

La hipótesis de independencia entre las dos variables se rechaza para valores de χ_0^2 superiores a χ^2 con un grado de libertad y un nivel de significancia α. Por ejemplo, los datos sobre la gripe de la tabla 1.1, los

Tabla 2.15: Resultados respiratorios.

	Resultado		
Tratamiento	Favorable	Desfavorable	Total
Placebo	16	48	64
Fármaco	40	20	60
Total	56	68	N=124

cuales se transcriben en la tabla 2.15, producen el siguiente valor de χ_0^2:

$$\chi_0^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{124(16\times 20 - 48\times 40)^2}{64\times 60\times 56\times 68}$$
$$= 21.7087$$

A un nivel de significancia de 5% el valor de χ^2 es 3.84 (tabla A.2). El valor calculado de $\chi_0^2 = 21.7087$ es mayor que 3.84, luego se puede advertir una asociación alta entre el consumo del medicamento para la gripe y el alivio.

2.5.2 La corrección por continuidad de Yates

Puesto que la distribución ji–cuadrado es una distribución continua, corresponde a la distribución aproximada (asintótica) para la distribución discreta de las frecuencias observadas. Yates (1934) sugirió una corrección que se incorpora directamente a la expresión (2.40), como se muestra en seguida:

$$\chi_0^2 = \frac{N(|ad-bc|-0.5N)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad (2.40a)$$

Esta se conoce como *el valor ji–cuadrado corregido por continuidad*. En (2.40a) el término $|ad - bc|$ significa el *valor absoluto* de $(ad - bc)$, y corresponde al valor numérico de la expresión sin el signo.

Para los datos de la tabla 2.15, el valor de la estadística ji–cuadrado corregida por continuidad es $\chi_0^2 = 20.059$, con la cual se deriva una decisión similar a la que se obtuvo con la estadística sin corregir.

2.5.3 Prueba de la probabilidad exacta de Fisher

Algunas veces los datos de una tabla de contingencia 2×2 incluyen frecuencias pequeñas o nulas. Por ejemplo, considere los datos de la tabla 2.16 relacionados con un estudio para la curación de infecciones severas. Un medicamento de prueba se compara frente a un tratamiento control para determinar si las proporciones de respuesta favorable son las mismas en los dos tratamientos. Los tamaños de muestra no

Tabla 2.16: Curación de infecciones severas.

	Resultado		
Tratamiento	Favorable	Desfavorable	Total
Medicamento	10	2	12
Control	2	4	6
Total	12	6	N=18

Fuente: Stokes, Davis y Koch (1997: 23)

son apropiados para aplicar la estadística ji–cuadrado. No obstante, si se pueden considerar las frecuencias marginales (12, 6, 12, 6) como cantidades fijas, entonces se puede asumir que los datos de las celdas tienen una distribución *hipergeométrica*, con la probabilidad que en la celda (i, j) haya n_{ij} observaciones dadas por

$$P(n_{ij}) = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{N!n_{11}!n_{12}!n_{21}!n_{22}!} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} \quad (2.41)$$

Esta probabilidad deberá calcularse para todas las posibles tablas de contingencia que permitan permanecer invariantes los totales marginales. En el ejemplo, los marginales por fila pueden fijarse mediante un proceso de asignación de tratamientos; es decir, los individuos son asignados

aleatoriamente al medicamento de interés o al de control. Los marginales por columna pueden ser considerados fijos a través de la hipótesis nula; es decir, hay 12 pacientes que reaccionan favorablemente y 6 que reaccionan desfavorablemente, sin considerar el tipo de tratamiento. Si los datos son el resultado de una muestra de conveniencia, se pueden fijar los totales marginales de acuerdo con la hipótesis nula de que los pacientes, indistintamente, pueden tener una respuesta favorable ante cualquiera de los dos tratamientos; caso en el que se ubicaría el mismo (o aproximadamente el mismo) número de pacientes en la columna de favorable y desfavorable.

El valor p es la probabilidad que los datos observados o datos más extremos ocurran bajo la hipótesis nula. Con la *prueba exacta de Fisher* se determina el valor p. Para la tabla 2.16, este se determina sumando las probabilidades de las tablas, que sin alterar los totales marginales, son tan probables o menos probables que la tabla disponible.

La tabla 2.17 incluye todas las posibles configuraciones de tablas y sus probabilidades asociadas. La probabilidad de obtener una tabla con las

Tabla 2.17: Probabilidades de las tablas 2×2.

Frecuencia por celda				
1-1	1-2	2-1	2-2	Probabilidad
12	0	0	6	0.0001
11	1	1	5	0.0039
10	2	2	4	0.0533
9	3	3	3	0.2370
8	4	4	2	0.4000
7	5	5	1	0.2560
6	6	6	0	0.0498

frecuencias contenidas en la tabla 2.16, es igual a:

$$P = \frac{12! \times 6! \times 12! \times 6!}{18! \times 10! \times 2! \times 2! \times 4!} = 0.0533$$

Para encontrar el valor p unilateral, se deben sumar las probabilidades tan pequeñas o más pequeñas que la calculada para la tabla 2.16, en la dirección determinada por la alternativa unilateral. Para este caso, serían las tablas para las cuales el tratamiento a investigar tendría más

favorabilidad en la respuesta; corresponden a las tablas de frecuencias (12, 0, 0, 6) y (11, 1, 1, 5), pues tienen respuestas favorables más altas (12 y 11) bajo el medicamento de interés. Así, se suma a la probabilidad de la tabla observada las probabilidades de las otras dos tablas, de donde resulta:

$$p = 0.0533 + 0.0039 + 0.0001 = 0.0573$$

Para encontrar el valor p bilateral, se suman todas las probabilidades que son tan pequeñas o más pequeñas que la probabilidad de la tabla observada. Este es igual a:

$$p = 0.0533 + 0.0039 + 0.0001 + 0.0498 = 0.1071$$

El interés, en general, se centra sobre el valor p bilateral. Note que cuando los marginales fila (o columna) son cercanamente iguales, el valor p para la prueba exacta bilateral de Fisher es aproximadamente el doble de una unilateral.

Dependiendo del nivel de significancia (α) asumido, se sacan las conclusiones sobre la dependencia entre los tratamientos y la condición de efectividad. Así, si la suma de estas probabilidades es más grande que un nivel de significancia α, no hay suficiente evidencia para afirmar acerca de la asociación entre las variables. Si, por el contrario, la suma de las probabilidades es menor que α, la hipótesis de independencia es puesta en duda.

El procedimiento FREQ del SAS calcula la estadística de Fisher para tablas de tamaño 2×2.

2.5.4 Prueba de McNemar para proporciones correlacionadas en tablas 2×2

El emparejamiento (capítulo 9) es una técnica que se emplea frecuentemente en estudios que utilizan el modelo de homogeneidad (Ato & López 1996, 37–38). Si se trata de un estudio experimental o un ensayo clínico controlado, se aparean los sujetos de acuerdo con alguna característica asociada con la respuesta y después se asignan al azar por pares a uno de dos (o más) grupos: *grupo tratamiento o experimental* (GE) y *grupo control* (GC). Si, por el contrario, se trata de estudios

observacionales comparativos (prospectivos o retrospectivos) o cuasiexperimentales, el apareamiento se aplica para incrementar la validez de las inferencias mediante el control de factores de confusión. Las variables de confusión más típicas son la edad, el género, el estrato social, el coeficiente de inteligencia, entre otras.

Los grupos resultantes del apareamiento no son independientes, sino correlacionados; en consecuencia, no es correcto aplicar las pruebas consideradas para este modelo. Como alternativa está la prueba de *McNemar*. A manera de presentación de esta prueba, se consideran los datos de la tabla 2.18 (Everitt 1994, 20), en los cuales se registra la presencia o ausencia de alguna característica A para muestras apareadas. El interés

Tabla 2.18: Frecuencias de muestras apareadas.

		Muestra I	
		A - ausente	A - presente
Muestra II	A - presente	a	b
	A - ausente	c	d

se centra en la diagonal principal, puesto que b se refiere a los pares (sujetos de la muestra I y II, respectivamente) donde los dos sujetos poseen el atributo A, y c a los pares de sujetos que no lo poseen. La comparación se dirige sobre las frecuencias a y d, donde a representa el número de pares que no poseen al atributo si provienen de la muestra I y lo poseen si proceden de la muestra II, d es el caso inverso. Bajo la hipótesis de que las dos muestras no difieren con respecto al atributo A, a y d deberían ser iguales, con valor esperado $(a+d)/2$. Al sustituir el valor observado y el esperado en la fórmula usual ji–cuadrado (2.15b), se obtiene

$$\chi^2_M = \frac{(a-d)^2}{a+d} \tag{2.42a}$$

Si se aplica corrección por continuidad, la expresión que resulta es:

$$\chi^2_M = \frac{(|a-d|-1)^2}{a+d} \tag{2.42b}$$

la cual corresponde a la fórmula de McNemar para verificar la asociación en una tabla 2×2 cuando las muestras son mezcladas. Bajo la hipótesis

de no diferencia entre las muestras mezcladas con relación al atributo A, la estadística (2.42b) tiene una distribución ji–cuadrado con un grado de libertad.

A continuación se ilustra el uso de esta estadística mediante los datos de (Everitt 1994, 20), los cuales se muestran en la tabla 2.19. Los datos corresponden a la evaluación que un psiquiatra desea hacer sobre el efecto del síntoma de despersonalización con relación al pronóstico de pacientes deprimidos. Para esto, un grupo de 23 pacientes con depresión endógena diagnosticados con síntomas de despersonalización (grupo experimental, GE) fue apareado de acuerdo con la edad, género, duración de la enfermedad y otros factores, con 23 pacientes con depresión endógena, quienes no presentaban el síndrome de despersonalización (grupo control, GC). Después de un periodo terapéutico, los 23 pares de pacientes fueron evaluados como recuperados o no recuperados. Las dos versiones

Tabla 2.19: Recuperación en pacientes depresivos.

		Pacientes despersonalizados		
		No recuperado	Recuperado	Total
Pacientes no despersonalizados	Recuperado	5	14	19
	No recuperado	2	2	4
	Total	7	16	23

de la estadística de McNemar, evaluadas sobre estos datos, son:

$$\chi^2_M = \frac{(a-d)^2}{(a+d)} = \frac{(5-2)^2}{5+2} = 1.2857$$
$$\chi^2_M = \frac{(|a-d|-1)^2}{(a+d)} = \frac{(|5-2|-1)^2}{5+2} = 0.5714$$

En los dos casos, estos valores son menores que el valor crítico para un $\alpha = 0.05$ de ji–cuadrado con un grado de libertad, es decir que, $\chi^2_{(0.05,1)} = 3.84$. Se puede concluir, de estos datos, que el síntoma de despersonalización no tiene efecto significativo sobre el pronóstico de pacientes deprimidos.

La prueba de McNemar también es aplicable a situaciones en las cuales los mismos individuos se observan en dos ocasiones. Por ejemplo, suponga que dos medicamentos A y B son usados para tratar la depresión, y que

se quieren comparar en términos de la presencia de náusea como efecto colateral. Los dos medicamentos son suministrados a los pacientes en dos ocasiones, al cabo de las cuales se registra la presencia o no de náusea. En este caso, como en el anterior, se trata de observaciones correlacionadas, pues los mismos sujetos reciben los medicamentos A y B. La tabla 2.20 muestra los datos resultantes de la investigación. Para contrastar la hipótesis de incidencia entre el tipo de medicamento

Tabla 2.20: Sujetos que muestran náusea con las drogas A y B.

		Droga A		
		No náusea	Náusea	Total
Droga B	Náusea	3	9	12
	No náusea	75	13	88
	Total	78	22	100

Fuente: (Everitt 1994, 21)

y la presencia de náusea en estos pacientes, se calcula la estadística de McNemar:

$$\chi^2_M = \frac{(|3-13|-1)^2}{3+13} = 5.0625$$

Este valor es superior a $\chi^2_{(0.05,1)} = 3.84$, luego se puede concluir, con un nivel de incertidumbre de 5%, que hay evidencia de que los dos medicamentos inciden en la náusea de manera diferente.

Observación

- Con el procedimiento FREQ del SAS, se obtiene el valor de la estadística de McNemar, sin corregir, y su valor p, escribiendo `AGREE` en la instrucción `TABLES` como sigue:

```
TABLES DROGA_B*DROGA_A/AGREE;
```

2.5.5 Riesgo relativo

Una diferencia entre dos proporciones de algún tamaño fijo puede tener más importancia cuando ambas proporciones están cerca de 0.0 o de

1.0 que cuando están en valores intermedios. Suponga que se quieren comparar dos medicamentos respecto a la proporción de individuos que tienen reacciones adversas cuando usan el medicamento. La diferencia entre, por ejemplo, 0.010 y 0.001 es igual a la de entre 0.410 y 0.401, nominalmente 0.009. La primera diferencia parece más notable, dado que con una medicina se tiene 10 veces más probabilidad de tener reacciones adversas que con relación a la otra. En tales casos, la razón de las proporciones es una medida descriptiva útil.

Suponga una tabla 2×2, donde las filas son las modalidades A y B y las columnas las modalidades E y E^c, y que $P_A(E)$ denota el riesgo[3] de que un individuo contraiga la enfermedad E[4], cuando está presente (expuesto a) la condición A, y que B es otra condición (factor de riesgo), de manera que $P_B(E)$ denota el riesgo que se corre de padecer la misma enfermedad bajo la condición B. En tablas 2×2, la razón

$$RR = \frac{P_A(E)}{P_B(E)} = \frac{P(E|A)}{P(E|B)} \tag{2.43}$$

expresa el *riesgo relativo* de padecer la enfermedad E cuando se está en la condición A respecto de cuando se está en la condición B. En otras palabras, sintetiza cuánto más probable es desarrollar la enfermedad si se está en el primer caso que si se está en el segundo.

Un caso particular de este concepto es aquel en que se analiza una enfermedad E; la condición A es haber estado expuesto a cierto factor y la condición B es la complementaria: no haber estado expuesto a él. En tal caso se dice que RR es el riesgo relativo inherente al factor y queda claro a qué se refiere.

Las proporciones 0.01 y 0.001 tienen un riesgo relativo de $0.01/0.001 = 10.0$, mientras que las proporciones 0.410 y 0.401 tienen un riesgo relativo de $0.410/0.401 = 1.02$. Un riesgo relativo de 1.00 ocurre cuando las probabilidades del evento son iguales bajo los dos factores.

Para el caso de dos grupos con proporciones muestrales p_1 y p_2, respecto a un evento, el *riesgo relativo muestral* es p_1/p_2.

[3]Probabilidad, aunque usualmente es una tasa de prevalencia o de incidencia.

[4]Además de una enfermedad, puede ser otro evento como tener un accidente, ser portador de un virus, etc. Por comodidad, con frecuencia se hablará de "enfermedad".

Los datos de la tabla 2.21 muestran la relación en el uso de la aspirina en infartos del miocardio (ataques cardiacos). Se trata de verificar si el consumo regular de la aspirina reduce la mortalidad por enfermedades cardiovasculares. Las personas que participaron del estudio, durante cinco años, consumieron diariamente una tableta de aspirina o un placebo. Estas personas no conocían el tipo de pastilla que se les administraba (estudio ciego). De la tabla 2.21, la probabilidad de padecer infartos

Tabla 2.21: Consumo de aspirina e infartos del miocardio.

	Infartos del miocardio		
Grupo	Sí	No	Total
Placebo	189	10845	11034
Aspirina	104	10933	11037
Total	293	21778	22071

Fuente: (Agresti 1996, 20).

del miocardio para los grupos placebo y aspirina son, respectivamente, $p_1 = 189/11034 = 0.0171$ y $p_2 = 104/11037 = 0.0094$. El riesgo relativo es $p_1/p_2 = 0.0171/0.0094 = 1.82$. De manera que la proporción de casos con infartos del miocardio es el 82% más alto en el grupo que toman el placebo.

Mediante el PROC FREQ del SAS se obtiene, además de la estimación del RR, un intervalo de confianza para este. Se trata entonces de observar si el intervalo contiene o no a 1.0. Si ambos extremos del intervalo son mayores que 1.0 se puede concluir que el evento E tiene más probabilidad de ocurrir bajo la condición A que bajo la B. Si los dos extremos son menores que 1.0, se puede concluir que es menos probable que ocurra E bajo el factor A que bajo el factor B. Si el intervalo contiene a 1.0, se puede concluir que la probabilidad de ocurrencia de E es aproximadamente la misma bajo los dos factores de riesgo. En este último punto se debe tener en cuenta que si bien la probabilidad de ocurrencia, bajo los dos factores, es aproximadamente igual, esto no permite descuidar el bajo o alto riesgo de ocurrencia del evento de interés.

De los datos del consumo de aspirina, el intervalo de confianza para el riesgo relativo es $(1.433, 2.306)$. Se puede concluir con un 95% de confianza que la proporción de infartos del miocardio en personas que toman el placebo está entre 1.433 y 2.306 veces de la proporción de infartos del miocardio de personas que consumen aspirina. De otra ma-

nera, el riesgo de infartos del miocardio es al menos el 43% más alto para el grupo placebo. Un intervalo de confianza para la diferencia de proporciones es $(0.005, 0.011)$, con el cual se afirmaría que las proporciones de los grupos difieren por una cantidad muy escasa, tanto que se podría afirmar que no difieren significativamente. Note que, en cambio, el riesgo relativo advierte que la diferencia puede ser importante.

Algunas veces puede ser de interés computar la razón de las proporciones referentes al evento complementario (para el caso no infarto); el riesgo relativo resulta entonces igual a $[1 - P_A(E)]/[1 - P_B(E)]$.

2.5.6 Razón de probabilidades (*odds*)

La traducción más aproximada del término sajón *odds* es "la ventaja", en términos de probabilidades, que un evento ocurra con relación a que no ocurra; es decir, es un número que expresa cuánto más probable es que se produzca un evento frente a que no se produzca.

Sea E el evento de interés, $P(E)$ la probabilidad que ocurra y $O(E)$ el *odds* que le corresponde, entonces se tiene:

$$O(E) = \frac{P(E)}{1 - P(E)} \tag{2.44}$$

Los *odds* son números no negativos, con valores mayores que 1.0 cuando es más probable que ocurra el evento (éxito) frente a que no ocurra (fracaso).

Por ejemplo, si se estima que el 75% de los pacientes que ingresan al pabellón de quemados de un hospital sobreviven, se dice que los *odds* de que un paciente de éstos sobreviva son 3, pues $O(E) = 0.75/0.25 = 3$; es decir hay el triple de probabilidad de que sobreviva contra que no. En ambientes deportivos también se habla de los *odds* que tiene un equipo de baloncesto antes de un juego; con ello se alude a cuántas veces es más probable que gane frente a que pierda.

Por tanto, conocidos los *odds*, se puede deducir la probabilidad asociada al evento. De esta manera, si los *odds* de un evento E son iguales a $O(E)$, entonces, por la ecuación (2.44), se tiene que

$$P(E) = \frac{O(E)}{1 + O(E)} \tag{2.45}$$

Por ejemplo, si se sabe que los *odds* de sobrevivir que tiene un paciente operado de cáncer pulmonar son 0.4, esto equivale a decir que la probabilidad de que ese hecho ocurra es $0.4/1.4 = 0.285$.

Ambas informaciones son equivalentes y expresan la misma noción, es decir, cuantifican qué tan probable es que algo ocurra (en particular, cuál es el riesgo de un acontecimiento).

Naturalmente, entre la probabilidad y los *odds* del evento correspondiente hay una relación directa: si la probabilidad aumenta los *odds* también lo hacen y, recíprocamente, si la probabilidad disminuye los *odds* también disminuyen. Si $P(E) = 0$, entonces $O(E)$ también es nulo; pero en la medida que $P(E)$ tiende a uno, $O(E)$ tiende a infinito.

Los *odds* son una manera equivalente (aunque distinta) de expresar la probabilidad de un evento, de la misma manera que el *riesgo relativo (RR)* expresa la razón entre dos probabilidades, tiene sentido considerar la *razón de los odds*.

Así, se define la *razón de los odds* como el cociente entre el *odds*, asociado a un suceso bajo cierta condición y, el *odds* que le corresponderia bajo otra condición. Por este camino se mide la misma noción que con el RR, corresponde a la razón de *odds* entre dos filas en una tabla 2×2, es decir:

$$RO = \frac{O(E)_F}{O(E)_{F^c}} = \frac{P_F(E)/(1 - P_F(E))}{P_{F^c}(E)/(1 - P_{F^c}(E))} \tag{2.46a}$$

donde F^c expresa la no exposición al factor de riesgo. Otra notación empleada para las probabilidades por celda en una tabla 2×2 es π_{ij}, con $i, j = 1, 2$, con esta notación, la razón de *odds* (2.46a) es equivalente a

$$RO = \frac{O(E)_F}{O(E)_{F^c}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{2.46b}$$

De aquí que a la razón de *odds* se le denomine también *razón de productos cruzados*, puesto que tanto el numerador como el denominador son iguales al producto de las probabilidades de celdas opuestas por la diagonal principal y secundaria, respectivamente.

Una última expresión para la razón de *odds* es la relacionada con los valores esperados por celda F_{ij}. Puesto que $F_{ij} = N\pi_{ij}$ entonces, $\pi_{ij} = F_{ij}/N$; así,

$$RO = \frac{F_{11}F_{22}}{F_{12}F_{21}} \tag{2.46c}$$

La razón de *odds* muestral es igual a la razón de los *odds* muestrales en las dos filas:

$$\widehat{RO} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{2.46d}$$

La razón de *odds* es un número no negativo. Cuando las variables fila y columna son independientes, entonces $P_F(E) = P_{F^c}(E)$; en consecuencia $O(E)_F = O(E)_{F^c}$ y $RO = 1$. El valor de $RO = 1.0$, correspondiente al caso en el que las variables fila y columna son independientes, sirve como referente para la comparación. La figura 2.4 ilustra las situaciones más importantes que se pueden presentar con el valor de RO en una tabla de contingencia 2×2.

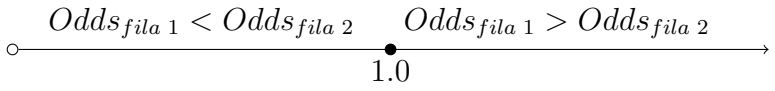

Figura 2.4: Valores de la razón de *odds* (RO).

Cuando $RO > 1$, los sujetos de la fila 1 tienen más probabilidad de "éxito" que los de la fila 2. Cuando $0 < RO < 1$, el "éxito"es menos probable en la fila 1 que en la fila 2. Valores de RO lejanos de 1.0, en cualquier dirección, señalan niveles de asociación fuerte entre las variables. Por ejemplo, un valor de RO de 4 señala más dependencia entre las variables que un valor de 2, y un valor de 0.1 indica más dependencia que un valor de 0.5. Si un valor de RO es el inverso de otro, estos representan el mismo nivel de asociación, pero en dirección opuesta; por ejemplo, si $RO = 0.25$, los *odds* de "éxito" en la fila 1 son 0.25 veces los *odds* de "éxito" de la fila 2, o equivalentemente, $1/0.25 = 4.0$ veces más alto en la fila 2 que en la 1. Cuando el orden de las filas (o las columnas) se invierte, el valor de RO es el recíproco del valor original. Otra característica de los *odds* es que no cambian al transponer la tabla, es decir, cuando las filas se hacen columnas y las columnas filas.

Para los datos sobre el consumo de aspirina e infartos del miocardio (tabla 2.21), para quienes toman el placebo, el *odds* de infartos del miocardio es estimado por $n_{11}/n_{12} = 189/10845 = 0.0174$, el cual significa que en este grupo (no consumidores de aspirina) hubo 17.4 infartos por cada 1.000 de "no" infartos. Para quienes toman aspirina, el *odds* estimado es igual a $104/10933 = 0.0095$, es decir, 9.5 infartos por cada 1.000 de no infartos.

La razón de *odds* muestral es igual a $\widehat{RO} = 0.0174/0.0095 = 1.832$. Así, el *odds* estimado de infartos del miocardio para quienes toman placebo es 1.832 veces el *odds* estimado para quienes toman aspirina.

A continuación se presentan las instrucciones SAS para el ejemplo de infartos del miocardio y el consumo de aspirina.

```
DATA   CARDIO;
INPUT GRUPO$ INFARTO$   FRECUE;
CARDS;
placebo    si   189   placebo    no   10845
aspirina   si   104   aspirina   no   10933
;
PROC FREQ;
WEIGHT FRECUE;
TABLES GRUPO*INFARTO /MEASURES;
RUN;
```

En la salida que se presenta a continuación, la primera línea corresponde a la razón de *odds RO* y un intervalo de confianza de 95% asociados a los infartos del miocardio de las personas que no consumieron aspirina, respecto a las que sí consumieron.

Estimates of the Relative Risk (Row1/Row2)

		95%	
Type of Study	Value	Confidence	Bounds
Case-Control	1.832	1.440	2.331
Cohort (Col1 Risk)	1.818	1.433	2.306
Cohort (Col2 Risk)	0.992	0.989	0.995

Sample Size = 22071

La segunda y tercera línea contienen el riesgo relativo (no consumir o consumir aspirina) de sufrir y no sufrir infarto del miocardio, respectivamente, junto con un intervalo de confianza del 95% para cada riesgo relativo.

2.5.7 Fracción etiológica

La *fracción etiológica* λ es una medida del impacto de una exposición sobre la población, a diferencia del riesgo relativo, el cual mide el impacto de la exposición sobre una subpoblación expuesta. La fracción etiológica se define como la proporción de casos enfermos atribuibles al factor de riesgo. Para los datos de la tabla 2.21, el interés apunta a la proporción de infartados atribuibles al placebo (expuestos) (aunque podría ser también al consumo de aspirina; no expuestos). Asuma que el tamaño, en un momento dado, de la población es N y que N_1 es el número de casos (infartados). Si p_1 y p_2 son la tasas de infartados expuestos y no expuestos en las respectivas subpoblaciones (no consumidores y consumidores de aspirina), entonces

$$\lambda = \frac{(N_1 - Np_2)}{N_1} \tag{2.47a}$$

expresa la proporción de casos de enfermedad entre los expuestos al determinante asociado, debidos a la exposición a dicho determinante.

Si la exposición no tiene efecto, el número de casos esperado es Np_2. Si la tasa de exposición poblacional es p_e (porcentaje de no consumidores de aspirina en la población), se tiene que

$$\lambda = \frac{p_e(RR - 1)}{1 + p_e(RR - 1)} \tag{2.47b}$$

donde RR es el riesgo relativo asociado a la exposición (no consumo de aspirina).

La fracción etiológica, como se muestra en (2.47b), está en relación directa con RR (es una función creciente de RR). Así:

Si $RR > 1$	$\lambda > 0$ (efecto "favorable")
Si $RR < 1$	$\lambda < 0$ (efecto "desfavorable")
Si $RR = 1$	$\lambda = 0$ (efecto "indiferente")

Esta expresión no es muy útil, toda vez que en la práctica no siempre se conoce p_e. Cuando los datos son presentados conforme a entonces una estimación de la fracción etiológica es

$$\widehat{\lambda} = \frac{a}{n_1}\left[1 - \frac{bc}{ad}\right] \tag{2.47c}$$

Expuestos	Casos	Controles
$+$	a	b
$-$	c	d
Tamaño de muestra	n_1	n_2

o también

$$\widehat{\lambda} = 1 - \frac{cn_2}{dn_1} \tag{2.47d}$$

Un intervalo del 95% de confianza para λ es dado por

$$1 - (1 - \widehat{\lambda}) \exp\left\{ \mp 1.96 \sqrt{\frac{a}{cn_1} + \frac{b}{dn_2}} \right\} \tag{2.47e}$$

Los datos de la tabla 2.21, sobre infartos del miocardio por el no consumo o consumo de aspirina, se rescriben en la tabla 2.22. Estos datos

Tabla 2.22: Consumo de aspirina e infartos del miocardio.

Exposición	Casos	Controles
Placebo	189	10845
Aspirina	104	10933
tamaño de muestra	$n_1 = 293$	$n_2 = 21778$

corresponden a $n_1 = 293$ (con $a = 189$, $c = 104$) y $n_2 = 21788$ (con $b = 10845$, $d = 10933$). De acuerdo con estos datos y la expresión (2.47d), la estadística $\widehat{\lambda}$ toma el valor

$$\begin{aligned} \widehat{\lambda} &= 1 - \frac{cn_2}{dn_1} = 1 - \frac{(104)(21788)}{(10933)(293)} \\ &= 1 - \frac{2265952}{3203369} \\ &= 1 - 0.7074 = 0.2926 \end{aligned}$$

Con una lectura literal del resultado anterior se puede afirmar que el no consumo de aspirina (placebo) favorece los infartos del miocardio; de otra manera, el consumo de aspirina disminuye el riesgo de infartos del miocardio.

Si un intervalo del 95% de confianza para la fracción etiológica, de (2.47e), es $[0.1741;\ 0,3941]$, significa que en esta población de no consumidores de aspirina los infartos del miocardio se aumentan entre el $17,41$ y el $39,41$ por ciento.

2.5.8 Prueba de Cochran–Mantel–Haenszel

Con frecuencia se quiere indagar sobre la relación entre dos variables dicotómicas (por ejemplo, enfermedad y exposición), la cual puede estar afectada por la presencia de una tercera variable; a esta última variable se le denomina *factor de confusión*: es decir, una variable asociada tanto a la enfermedad como al riesgo (sección 2.5.4).

Se trata entonces de presentar una prueba de independencia condicional y una prueba de homogeneidad de asociación de las variables contenidas en tablas de la forma $2 \times 2 \times k$. Los análisis de *asociación condicional* son bastante útiles en la mayoría de las aplicaciones con datos multivariados. Un ejemplo es la indagación sobre la posible asociación entre las variables ocurrencia de cáncer pulmonar y el tabaquismo, teniendo como tercera variable la región de residencia (tres regiones diferentes). Los individuos de una región pueden tener características muy diferentes (estrato socioeconómico, prácticas culturales, etc.) respecto de los individuos no fumadores, de manera que esta variable puede verse asociada tanto a la enfermedad como al tabaquismo. Específicamente, se quiere saber si en cada región están asociados el hábito de fumar y el padecimiento de cáncer pulmonar.

Se podrían desarrollar análisis para verificar la independencia entre el cáncer y el tabaquismo en cada uno de los tres sitios. Esta estrategia correspondería a un análisis estratificado; donde los estratos (sitios, centros de salud, género, etc.) son determinados por los niveles de las variables explicatorias (uno por cada combinación). Aunque podrían hacerse los análisis parciales para cada tabla (una por cada nivel de la variable de confusión), la idea es investigar la asociación global.

Asumiendo que la variable de confusión (sitio) no tiene efecto modificador sobre la relación entre la enfermedad y el factor de riesgo, se mezclan los datos para tomar una decisión conjunta.

La idea de la metodología de Mantel–Haenszel es remover la influencia de los factores de confusión sobre las variables explicativas, para suministrar

una prueba que, mediante la comparación de individuos homogéneos dentro de cada uno de los niveles (estratos) de la tercera variable, detecte asociación entre las variables de interés.

Para tablas de la forma $2 \times 2 \times k$, la hipótesis nula es que X y Y son condicionalmente independientes, dada la variable Z. El proceso puede ser resumido de la siguiente forma:

(i) Se forman tablas 2×2, una en cada uno de los k niveles de la variable de confusión.

(ii) En cada uno de los h niveles ($h = 1, 2, \ldots, k$) de la variable de confusión se tiene una tabla de la forma:

	Padece de cáncer		
Fuma	Sí	No	Total
Sí	n_{11h}	n_{12h}	$n_{1.h}$
No	n_{21h}	n_{22h}	$n_{2.h}$
Total	$n_{.1h}$	$n_{.2h}$	n_h

Bajo la hipótesis nula y con los totales marginales fijos, la frecuencia "n_{11h}" de la celda $(1, 1)$ tiene distribución hipergeométrica con media y varianza

$$\mathrm{E}(n_{11h}|H_0) = \mu_{11h} = \frac{n_{1.h}n_{.1h}}{n_h}$$
$$\mathrm{var}(n_{11h}|H_0) = \sigma^2_{11h} = \frac{n_{1.h}n_{2.h}n_{.1h}n_{.2h}}{n_h^2(n_h - 1)}$$

La estadística de Cochran-Mantel-Haenszel resume la información de las k tablas parciales mediante la siguiente expresión:

$$CMH = \frac{\left[\sum_{h=1}^{k}(n_{11h} - \mu_{11h})\right]^2}{\sigma^2_{11h}} = \frac{\left[\sum_{h=1}^{k}(n_{11h} - \frac{n_{1.h}n_{.1h}}{n_h})\right]^2}{\sum_{h=1}^{k}\frac{n_{1.h}n_{2.h}n_{.1h}n_{.2h}}{n_h^2(n_h-1)}} \tag{2.48}$$

La estadística CMH es útil para detectar asociación a través de k *estratos* que tienen una tendencia fuerte a mostrar el mismo patrón de

dependencia entre las variables explicativas. Como se observa en la expresión (2.48), la estadística CMH es un *promedio de asociación parcial*. CMH puede fallar en la detección de asociación cuando las dependencias sean opuestas con una magnitud similar.

Para valores muestrales grandes, la estadística CMH tiene una distribución ji–cuadrado con 1 grado de libertad. Esta estadística no es apropiada cuando la asociación entre las variables X y Y varían notoriamente entre las tablas parciales (en cada una de las k tablas). La estadística CMH combina información a través de las tablas parciales (estratos). Cuando la asociación en cada tabla parcial es similar, la prueba es más potente que en cada una de las tablas por separado. No es conveniente combinar resultados de tablas parciales 2×2, puesto que se puede incurrir en la paradoja de Simpson (esta se explica en la próxima sección).

Como se asume que la variable de confusión no es un efecto modificador, es decir que la razón de *odds* es constante a través de sus niveles, la razón de *odds* es estimada por $n_{11h}n_{22h}/n_{12h}n_{21h}$; el procedimiento de Cochran-Mantel-Haenszel mezcla los datos para producir el estimador

$$RO_{CMH} = \frac{\frac{\sum_{h=1}^{k} n_{11h}n_{22h}}{n_h}}{\frac{\sum_{h=1}^{k} n_{12h}n_{21h}}{n_h}} \tag{2.49}$$

El procedimiento FREQ del SAS computa la estadística CMH, la razón de *odds* del mismo tipo y un intervalo de confianza para esta razón, que naturalmente supone la estimación de un error estándar asociado a RO_{CMH}.

La tabla 2.23 contiene datos sobre un estudio de enfermedades respiratorias de un grupo de personas que tenían la misma sintomatología. A cada una de las personas se les asignó uno de dos tratamientos: a un grupo se le aplicó un medicamento, y al otro un placebo. El estudio incluyó pacientes de tres centros médicos.

Con los datos de la tabla 2.23, el investigador está interesado en verificar si hay diferencias en las tasas de mejoría entre el medicamento y el placebo. Los pacientes en los tres centros fueron asignados aleatoriamente a uno de los dos tratamientos[5]. De acuerdo con (2.48), la

[5]El análisis estadístico que combina información de varios estudios se llama

Tabla 2.23: Mejoría en enfermedades respiratorias.

Centro Méd.	Tratamiento	Mejoría Sí	Mejoría No	Total
1	Medicamento	29	16	45
	Placebo	14	31	45
Total		43	47	90
2	Medicamento	37	8	45
	Placebo	24	21	45
Total		61	29	90
3	Medicamento	30	15	45
	Placebo	23	22	45
Total		53	37	90

estadística CMH se calcula como se muestra a continuación.

$$
\begin{aligned}
CMH &= \frac{\left[\sum_{h=1}^{k}(n_{11h} - \frac{n_{1.h}n_{.1h}}{n_h})\right]^2}{\sum_{h=1}^{k}\frac{n_{1.h}n_{2.h}n_{.1h}n_{.2h}}{n_h^2(n_h-1)}} \\
&= \frac{\left[(29 - \frac{45\times43}{90}) + (37 - \frac{45\times61}{90}) + (30 - \frac{45\times53}{90})\right]^2}{\left(\frac{45\times45\times43\times47}{90^2(90-1)}\right) + \left(\frac{45\times45\times61\times29}{90^2(90-1)}\right) + \left(\frac{45\times45\times53\times37}{90^2(90-1)}\right)} \\
&= \frac{[17.5]^2}{16.5084} \\
&= 18.9576
\end{aligned}
$$

El valor de $CMH = 18.9576$ es altamente significativo respecto al percentil 99 de una ji–cuadrado con un grado de libertad (de la tabla A.2, este valor es 6.63). De manera que se puede afirmar la existencia de una asociación fuerte entre mejoría (respuesta) y tratamiento, ajustada por el centro de salud. El tratamiento bajo prueba mostró una favorabilidad significativamente superior a la respuesta que el placebo.

El estimador de la razón de *odds* tipo CMH, de acuerdo con la expresión

metaanálisis.

(2.49), es

$$\begin{aligned} RO_{CMH} &= \frac{\frac{\sum_{h=1}^{k} n_{11h}n_{22h}}{n_h}}{\frac{\sum_{h=1}^{k} n_{12h}n_{21h}}{n_h}} \\ &= \frac{\left(\frac{29\times 31}{90}\right) + \left(\frac{37\times 21}{90}\right) + \left(\frac{30\times 22}{90}\right)}{\left(\frac{16\times 14}{90}\right) + \left(\frac{8\times 24}{90}\right) + \left(\frac{15\times 23}{90}\right)} \\ &= \frac{2336}{761} \\ &= 3.0696 \end{aligned}$$

Este valor corrobora, por lo menos de manera descriptiva, que el medicamento es tres veces más favorable a la mejoría que el placebo.

En el siguiente cuadro se indican las instrucciones con las cuales se hacen los cálculos de estas estadísticas.

```
DATA respira;
INPUT centrom$ trata$  resp$ conteo @@;
CARDS;
1 Medicam si 29 1 Medicam no 16
1 Placebo si 14 1 Placebo no 31
2 Medicam si 37 2 Medicam no  8
2 Placebo si 24 2 Placebo no 21
3 Medicam si 30 3 Medicam no 15
3 Placebo si 23 3 Placebo no 22
;
PROC  FREQ  ORDER=DATA;
WEIGHT  conteo;
TABLES  centrom*trata*resp / CHISQ MEASURES CMH;
RUN;
```

Una generalización de la estadística CMH para el caso de tablas $a \times b \times k$ se encuentra en Stokes, Davis & Koch (1997, 106).

2.6 Tablas multidimensionales

Los análisis realizados hasta aquí se han centrado sobre tablas de contingencia bidimensionales, es decir, tablas que cruzan las modalidades de dos variables categóricas. Como señala Everitt (1994, 60) el análisis de tablas con dimensión superior a dos tiene problemas conceptuales totalmente nuevos con relación a los que se observaron para tablas de doble entrada. La extensión hacia tablas de dimensión superior, aunque tiene algunos elementos para el análisis y la interpretación más complejos, no supone problemas conceptuales fuertes.

2.6.1 Notación para tablas multidimensionales

La notación es una generalización de la empleada en tablas bidimensionales. Con el objeto de concretar algunos conceptos se ilustrará su presentación a través de tablas de dimensión tres. Así, para tres variables A, B y C, n_{ijk} representa la frecuencia observada para la celda determinada por la modalidad i de la variable A, la modalidad j de la variable B y la modalidad k de la variable C. De manera análoga, E_{ijk} representa la frecuencia esperada para la celda $i-j-k$. Se supone que el número de modalidades de las variables A, B y C son a, b y c, respectivamente. La tabla 2.24 representa las frecuencias por celda y las marginales en una tabla de tamaño $a \times b \times c$. La presentación de las demás frecuencias marginales no es fácil en la tabla 2.24; a cambio, se presentan a continuación las expresiones que ilustran la manera de obtenerlas.

$n_{i..} = \sum_{j=1}^{b} \sum_{k=1}^{c} n_{ijk};\ i = 1, 2, \cdots, a$ (Marginales de A)

$n_{.j.} = \sum_{i=1}^{a} \sum_{k=1}^{c} n_{ijk};\ j = 1, 2, \cdots, b$ (Marginales de B)

$n_{..k} = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ijk};\ k = 1, 2, \cdots, c$ (Marginales de C)

$n_{ij.} = \sum_{k=1}^{c} n_{ijk};\ i = 1, 2, \cdots, a;\ j = 1, 2, \cdots, b$ (Marginales de $A \times B$)

$n_{i.k} = \sum_{j=1}^{b} n_{ijk};\ i = 1, 2, \cdots, a;\ k = 1, 2, \cdots, c$ (Marginales de $A \times C$)

$n_{.jk} = \sum_{i=1}^{a} n_{ijk};\ j = 1, 2, \cdots, b;\ k = 1, 2, \cdots, c$ (Marginales de $B \times C$)

Tabla 2.24: Tabla de contingencia tridimensional.

		Variable C					
Variable A	Variable B	1	$\cdots$	k	$\cdots$	c	Totales
1	1	n_{111}	$\cdots$	n_{11k}	$\cdots$	n_{11c}	$n_{11.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	b	n_{1b1}	$\cdots$	n_{1bk}	$\cdots$	n_{1bc}	$n_{1b.}$
2	1	n_{211}	$\cdots$	n_{21k}	$\cdots$	n_{21c}	$n_{21.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
2	b	n_{2b1}	$\cdots$	n_{2bk}	$\cdots$	n_{2bc}	$n_{2b.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a	1	n_{a11}	$\cdots$	n_{a1k}	$\cdots$	n_{a1c}	$n_{a1.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a	b	n_{ab1}	$\cdots$	n_{abk}	$\cdots$	n_{abc}	$n_{ab.}$
Total	Total	$n_{..1}$	$\cdots$	$n_{..k}$	$\cdots$	$n_{..c}$	$n_{...} = N$

$$N = n_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk};\ i = 1, 2, \cdots, a;\ j = 1, 2, \cdots, b;$$
$$k = 1, 2, \cdots, c \text{ (Total)} \quad (2.50)$$

2.6.2 Pruebas de independencia de las variables en una tabla a tres vías

La hipótesis de independencia mutua de las variables en una tabla de contingencia, la cual es una extensión de la igualdad (2.11), se formula como sigue:

$$H_0 :\ \pi_{ijk} = (\pi_{i..})(\pi_{.j.})(\pi_{..k}) \quad (2.51)$$

donde π_{ijk} representa la probabilidad de que una observación "caiga" en la celda $i - j - k$ de la tabla, y $\pi_{i..}$, $\pi_{.j.}$ y $\pi_{..k}$ son las probabilidades marginales de las variables A, B y C, respectivamente. Para probar esta hipótesis, como en la sección (2.3), se obtienen las frecuencias esperadas estimadas E_{ijk} cuando H_0 es cierta y se determina la distancia, tipo ji–cuadrado, entre las observaciones y los valores esperados estimados. Finalmente, se compara la distancia anterior con el valor de significancia de la estadística ji–cuadrado respectiva.

El valor de los E_{ijk}, bajo la independencia de las tres variables, es

$$E_{ijk} = N(\widehat{\pi}_{i..})(\widehat{\pi}_{.j.})(\widehat{\pi}_{..k}) = \frac{n_{i..}n_{.j.}n_{..k}}{N^2} \tag{2.52}$$

donde $\widehat{\pi}_{i..}$, $\widehat{\pi}_{.j.}$ y $\widehat{\pi}_{..k}$ son estimadores de las probabilidades correspondientes.

La estadística de prueba es:

$$\chi_0^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk} - E_{ijk})^2}{E_{ijk}} \tag{2.53}$$

la cual tiene como grados de libertad ($g.l.$) a la cantidad

$$g.l. = abc - a - b - c + 2 \tag{2.54}$$

Se rechaza, con un nivel de significancia α, la hipótesis H_0 de independencia entre las tres variables si $\chi_0^2 > \chi_{g.l.,\alpha}^2$.

El rechazo de la hipótesis de independencia permite concluir que hay al menos un par de variables que son dependientes; de manera que el paso siguiente debe ser la búsqueda de las variables que tienen asociación y las que no. A veces se prefiere obviar esta prueba realizando pruebas para tablas bidimensionales parciales o condicionales, como se muestra a continuación.

2.6.3 Paradoja de Simpson

Los resultados de una asociación marginal pueden tener dirección diferente de las asociaciones condicionales, esto se conoce con el nombre de *paradoja de Simpson*. De otra manera, la relación entre las variables A y B puede ser distinta, dependiendo del control que se haga o no sobre una tercera variable C. Mediante unos casos se ilustra esta situación.

Considere un estudio, idealizado, que relaciona las enfermedades cardiacas (A), tabaquismo (B) y la edad (C). El análisis parcial (fijando el grupo de edad) del tabaquismo frente a las enfermedades cardiacas se puede desarrollar desagregando la tabla 2.25 conforme se muestra en las tablas 2.26 y 2.27.

Una primera alternativa consiste en tomar secciones transversales de la tabla 2.25 para formar tablas bidimensionales. Esto implica que se

Tabla 2.25: Enfermedades cardiacas por tabaquismo y edad.

	25 a 45 años		*mayor de 45 años*	
Enf. Cardiaca	Fuma	No fuma	Fuma	No fuma
Enfermo	143	65	107	10
No enfermo	78	26	39	52

tengan valores constantes para la tabla excluida, la cual pasa a ser una variable *controlada*. Las tablas resultantes son *tablas parciales* y el tipo de asociación a verificar es la *asociación parcial* entre dos variables, dado un nivel de la tercera. Para la información de la tabla 2.25, se trata de verificar la dependencia entre el tabaquismo y las enfermedades cardiacas en cada uno de los dos grupos de edad ($C = 1$ y $C = 2$). Las tablas de contingencia parcial se muestran como las tablas (2.26) y (2.27).

Tabla 2.26: Edad entre 25 y 45 años.

Enf. card.	Fuma	No fuma
Enf.	143	65
No enf.	78	26

Tabla 2.27: Edad superior a 45 años.

Enf. card.	Fuma	No fuma
Enf.	107	10
No enf.	39	52

La asociación entre el tabaquismo y las enfermedades cardiacas en el grupo de edad 25 a 45 años (tabla 2.26) muestra una estadística ji–cuadrado parcial de $\chi^2_{AB|C=1} = 1.311$, con un valor p de 0.252; es decir, de estos datos se puede afirmar que no hay una asociación fuerte entre el tabaquismo y las enfermedades cardiacas en personas cuya edad está entre 25 y 45 años. Mientras que la tabla 2.27 revela que existe una alta dependencia entre el tabaquismo y las enfermedades cardiacas en personas cuya edad es superior a los 45 años, pues $\chi^2_{AB|C=2} = 57.776$ y $p = 0.001$.

Una segunda alternativa consiste en tomar de una tabla tridimensional $a \times b \times c$ las tablas bidimensionales $a \times b$, $a \times c$ y $b \times c$; estas se denominan *tablas marginales* porque se obtienen mediante combinaciones de

marginales de la tabla tridimensional. La asociación que se verifica en una tabla marginal es la *asociación marginal* entre las dos variables. La variable que se excluye se ignora porque las frecuencias de sus categorías se suman.

Colapsando las tablas anteriores 2×2, correspondientes a los dos grupos de edad, se obtiene la tabla 2.28. Los resultados de la estadística ji–

Tabla 2.28: Tabaquismo y enfermedades cardiacas.

Enf. card.	Fuma	No fuma
Enf.	250	75
No enf.	117	78

cuadrado muestran que existe una alta asociación entre el tabaquismo y las enfermedades cardiacas, pues $\chi^2_{AB} = 16.808$ y $p = 0.001$.

Otro ejemplo, se muestra en la tabla 2.29 (Peña 1998, 398-399), en la cual se presenta la proporción de admitidos en una universidad clasificados por género. Si se supone homogeneidad y que los 4.000 estudiantes son una muestra aleatoria de la población de estudiantes pasados y futuros, se concluiría que hay una diferencia significativa en la admisión a favor de las mujeres. La tabla 2.30 presenta estos datos desglosados

Tabla 2.29: Admisiones a una universidad por género.

Género	Solicitudes	Admisiones	Proporción
Mujeres	2.000	1.136	56.80
Hombres	2.000	955	47.75

por facultades (letras, ingeniería y economía). Se observa que las tres facultades muestran discriminación a favor de los hombres, pues los porcentajes para estos son superiores. Por tanto, las conclusiones de los datos divididos en subpoblaciones más homogéneas (modalidades de la variable facultad) son opuestas a las de los datos agregados. Los ejemplos anteriores muestran un hecho significativo en el análisis de tablas de contingencia multidimensionales, esto es, que las asociaciones marginales entre variables pueden ser diferentes de sus correspondientes asociaciones parciales. Esta aparente contradicción, como la que se muestra entre las cinco últimas tablas, se conoce con el nombre de paradoja de *Simpson*. La paradoja se presenta como consecuencia de unir una o

Tabla 2.30: Admisiones por facultad y género.

Facultad	Género	Solicitudes	Admisiones	Proporción(%)
Letras	Mujeres	800	560	70
	Hombres	300	225	75
Ingeniería	Mujeres	200	36	18
	Hombres	700	140	20
Economía	Mujeres	1000	540	54
	Hombres	1000	590	59

más variables (en los casos anteriores, los grupos de edad y de facultad, respectivamente), pues resulta una población compuesta de poblaciones (modalidades de la tercera variable) no igualmente ponderadas.

La explicación de la paradoja es la siguiente: suponga una población con N elementos, dividida en k subpoblaciones de $N_1,\ N_2, \cdots, N_k$ elementos respectivamente (con $N = N_1 + N_2 + \cdots + N_k$). Sean $n_1,\ n_2, \cdots, n_k$ el número de elementos de cada subpoblación con la característica que se desea estudiar. Entonces, la proporción total de elementos con dicha característica es

$$p = \frac{n_1 + n_2 + \cdots + n_k}{N_1 + N_2 + \cdots + N_k} = \frac{\sum_i n_i}{N}$$

y sea p_i la probabilidad en cada subpoblación y $f_i = n_i/N$ su frecuencia relativa en el total, se puede escribir entonces

$$p = \sum_i f_i p_i$$

la cual indica que la probabilidad total es una media ponderada de las probabilidades parciales. Si se comparan dos sucesos A y B, es posible que p_A sea mayor que p_B en todas las subpoblaciones, pero que en la población general ocurra lo contrario, esto es, nuevamente, *la paradoja de Simpson*.

La paradoja no significa en modo alguno que haya análisis incorrectos, más bien advierte sobre la inconveniencia de trasladar las conclusiones que se observan respecto a la relación entre dos variables en una tabla, a las que se observarían, sobre las mismas variables, bajo algún valor de una tercera variable. El comentario también es válido para el caso recíproco.

No obstante existe una regla que sugiere cómo colapsar variables en una tabla de contingencia multidimensional para realizar un análisis de alguna subtabla (Bishop, Fienberg & Holland 1975, 47). La regla establece que para un conjunto de k variables categóricas que conforman una tabla multidimensional, se deben dividir las variables en tres grupos mutuamente excluyentes, así:

- Un primer grupo, G_I, compuesto por la(s) variable(s) que se desea(n) colapsar.
- Un segundo grupo, G_{II}, compuesto por la(s) variable(s) que es(son) independientes del primer grupo G_I.
- Un tercer grupo, G_{III}, compuesto por las restantes variables.

La regla señala que el grupo I de las variables es colapsable respecto al II, pero no respecto al III. Además, implica que los marginales que incluyan variables del grupo II no cambiarán cuando la tabla se colapse sobre una o más categorías de la(s) variable(s) del grupo I.

2.7 Tamaño de muestra

Esta es una de las inquietudes frecuentes en el diseño de una encuesta o en ensayos clínicos. De acuerdo con los propósitos del estudio, el tamaño de muestra puede ser aproximado por dos vías:

1. El control del ancho de un intervalo de confianza para el parámetro de interés.
2. El control sobre el riesgo de cometer un error tipo II (no rechazar una hipótesis que debiera rechazarse).

Suponga que el objetivo de un estudio es estimar una proporción π de sujetos de una población que cumplen cierta condición o poseen determinado rasgo. Un intervalo con una confianza del $(1-\alpha)\times 100\%$ para π (sección 1.5.3) viene dado por

$$p \mp Z_{\alpha/2}\sqrt{p(1-p)/n}$$

donde p es la proporción muestral. El ancho de este intervalo es igual a

$$2Z_{\alpha/2}\sqrt{p(1-p)/n}$$

De manera que si el estudio no es planificado adecuadamente, habrá la posibilidad real de que el intervalo resultante sea demasiado ancho para que pueda ser usado por el investigador. En tal caso, el investigador propone un tope (o una cota) para que la diferencia entre el verdadero valor del parámetro π y el valor muestral p no exceda a cierta cantidad d. A d se le denomina *error de estimación*, pues corresponde a la diferencia entre el parámetro π (lo real) y p (lo observado); es decir: $d = |\pi - p|$. En consecuencia, este objetivo es equivalente a hacer que la cantidad $Z_{\alpha/2}\sqrt{p(1-p)/n}$ no sea superior a d, esto es,

$$Z_{\alpha/2}\sqrt{p(1-p)/n} \leq d$$

Así, el mínimo tamaño de muestra requerido (despejando n de esta última expresión) es

$$n = \frac{Z^2_{\alpha/2}p(1-p)}{d^2} \tag{2.55}$$

Este tamaño de muestra es afectado por tres factores:

1. El grado de confianza $(1-\alpha)$;

2. El máximo error de estimación, d, determinado por el equipo de investigación.

3. La proporción p por sí misma.

El numeral 3 no deja de meternos en un círculo vicioso: todo este andamiaje tiene como finalidad conocer π, pero para conocerlo se debe utilizar p, que precisamente ¡se calculará una vez se obtenga la muestra! Algunas maneras de salirle al paso a esta incoherencia[6] son: mediante un estudio "piloto" o de una investigación similar realizada anteriormente (en la misma población o en una semejante); asumir $p = 0.5$, valor que maximiza la varianza de la respectiva variable dicotómica.

[6]Una discusión más amplia al respecto se encuentra en Silva (1995, Cap. 11).

Así, el tamaño de muestra mínimo requerido en un estudio para estimar una proporción π con una confiabilidad del 95% y un error de estimación d, en caso de carecer del conocimiento previo o colateral de p, es

$$n = \frac{(1.96^2)(0.25)}{d^2}$$

Considere ahora el problema donde se quiere diseñar un estudio para comparar dos proporciones asociadas a la respuesta ante dos tratamientos. Por ejemplo, una nueva vacuna que quiere probarse, para el estudio los individuos son asignados aleatoriamente en uno de dos grupos de igual tamaño: un grupo 1 de control (no inmunizado), y un grupo 2 experimental (inmunizado). Los sujetos de ambos grupos serán inoculados con cierto tipo de bacteria para comparar las tasas de infección. El juego de hipótesis a verificar es

$$H_0 : \ \pi_1 = \pi_2 \text{ frente a } H_1 : \ \pi_1 \neq \pi_2$$

Una inquietud del equipo investigador tiene que ver con el tamaño de muestra adecuado para el desarrollo de este estudio.

Suponga que es importante detectar una tasa de reducción de la infección igual a

$$\delta = \pi_1 - \pi_2$$

Se desea hallar el tamaño de muestra mínimo n que debe tomarse en cada grupo (el mismo para ambos) de modo que la prueba sea capaz de detectar como significativa (no atribuible al azar) una diferencia mínima δ entre π_1 y π_2. La fórmula es

$$n = \frac{\left[Z_{1-\alpha/2}\sqrt{2\pi^*(1-\pi^*)} + Z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}\right]^2}{(\pi_1-\pi_2)^2} \quad (2.56)$$

donde α y β representan, de acuerdo con la sección 1.5.4, las probabilidades máximas admisibles de cometer, respectivamente, los errores tipo I (probabilidad de rechazar indebidamente la hipótesis nula) y de tipo II (el complemento de la potencia de la prueba) y $\pi^* = \frac{\pi_1+\pi_2}{2}$.

A manera de ilustración, retome el ejemplo de la vacuna. En este estudio se quiere determinar el número de individuos que se deben asignar a uno de dos grupos del mismo tamaño: un grupo control (no inmunizados) y un grupo experimental (inmunizados). Con base en conocimientos previos adquiridos desde un estudio piloto, se pueden establecer los siguientes supuestos:

1. La infección en el grupo control (cuando es inoculado con cierto tipo de bacteria) está cerca del 50%; es decir, $\pi_2 = 0.50$.

2. Alrededor del 80% del grupo experimental se espera que desarrolle anticuerpos adecuados. Si los anticuerpos son inadecuados, entonces la tasa de infección es casi igual a la del grupo control, pero si un sujeto del grupo experimental tiene anticuerpos adecuados, entonces se espera que la vacuna tenga cerca del 85% de efectividad (que corresponde a una tasa de infección del 15% contra la bacteria inoculada).

Poniendo estos supuestos conjuntamente, se obtiene un valor de π_1 equivalente a:

$$\begin{aligned}\pi_1 &= (0.80)(0.15) + (0.20)(0.50) \\ &= 0.22\end{aligned}$$

Además, suponga que se decide considerar $\alpha = 0.05$ y una potencia cerca del 90% (es decir, $\beta = 0.10$). En otras palabras,

$$\begin{aligned}Z_{1-\alpha/2} &= 1.96 \\ Z_{1-\beta} &= 1.28\end{aligned}$$

Así, el tamaño de muestra es

$$\begin{aligned}n &= \frac{\left[Z_{1-\alpha/2}\sqrt{2\pi^*(1-\pi^*)} + Z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}\right]^2}{(\pi_1-\pi_2)^2} \\ &= \frac{\left[1.96\sqrt{2(0.36)(0.64)} + 1.28\sqrt{(0.50)(0.50)+(0.22)(0.78)}\right]^2}{(0.22-0.50)^2} \\ &\approx 60\end{aligned}$$

De esta forma, cada grupo tendrá 30 individuos.

2.8 Procesamiento de datos con R

Se introduce la tabla 2.1 en la memoria como una matriz, después se asignan los nombres de filas y columnas, luego se convierte a un objeto de la clase tabla y por último se agregan las marginales.

```
t2.1<-matrix(c(75,60,20,15,35,70,30,25,40,50,40,40),4)

dimnames(t2.1)<-list(Estrato=c("Bajo","Medio","Alto",
                   "Muy Alto"),
                   Opinión=c("Bueno","Regular","Malo"))
t2.1<-as.table(t2.1)
tabla2.1<-addmargins(t2.1)
```

Se obtiene la tabla 2.3.

```
prop.table(t2.1)*100
addmargins(prop.table(t2.1)*100)
#Segunda tabla adicional.
prop.table(t2.1,1)*100
# Tercera tabla adicional.
prop.table(t2.1,2)*100
```

Gráficos de las distribuciones de frecuencias, figura 2.1.

```
par(mfrow=c(1,2))
# gráfico 2.1 a)
y<-prop.table(t2.1)
barplot(t(y),beside=T,xlab="(a) Estrato vs Opinión",
col="white", ylim=c(0,0.20), las=1,ylab="Porcentaje")

# gráfico 2.1 b)
y<-prop.table(t2.1)
barplot(y,beside=T,xlab="(b) Opinión vs estrato",
col="white",ylim=c(0,0.20),las=1,ylab="Porcentaje")
```

Gráficos de los perfiles fila, figura 2.2.

```
x<-prop.table(t2.1,1)
barplot(t(x),beside=T,xlab="Opinión",density =rep(20,4),
col="black",angle=c(45,90,135),las=1,ylab="Porcentaje")
legend("top",legend=colnames(x),fill="black",
 angle=c(45,90,135),density =rep(40,3),horiz=F,bty="n")
```

Prueba ji–cuadrado.

```
print(pji<-chisq.test(t2.1, correct=FALSE))
# frecuencias esperadas, tabla 2.7
pji$expected
# componentes de la ji--cuadrado
pji$residuals^2
```

Estadístico G^2.

```
Eij<-pji$expected
nij<-t2.1
G2<-2*sum( nij*log(nij/Eij))
G2
# p--valor
pchisq(G2,6,lower.tail=F)
# valor tabulado
qchisq(0.05,6,lower.tail=F)
```

Medidas de asociación.

```
# coeficiente de contingencia
Xo2<-pji$statistic
n<-sum(t2.1)
C<-sqrt(Xo2/(Xo2+n))
# coeficiente  phi
phi<-sqrt(Xo2/n)
# coeficiente V de Cramer
k<-min(nrow(t2.1)-1,ncol(t2.1)-1)
V<-sqrt(Xo2/(n*k))
```

Otra forma de obtener los estadísticos χ^2, G^2 y las medidas de asociación es usar la librería `vcd`[7], como se hace a continuación:

```
require(vcd) # si la librería no está instalada
             # regresa un FALSE
assocstats(t2.1)
```

Coeficientes lambda asimétricos (sección 2.4.6)

[7] Visualizing Categorical Data.

```
pt2.1<-prop.table(t2.1)
sipijmax<-sum(apply(pt2.1,2,max))
pi.max<-max( margin.table(pt2.1,1) )

# lambda Y|X
(sipijmax-pi.max)/(1-pi.max)

# lambda X|Y
sjpijmax<-sum(apply(pt2.1,1,max))
pj.max<-max( margin.table(pt2.1,2) )
(sjpijmax-pj.max)/(1-pj.max)
```

Coeficiente lambda simétrico.

```
(sipijmax+sjpijmax-pi.max-pj.max)/(2-pi.max-pj.max)
```

Estadística gama. Las siguentes funciones, tomadas de la página oficial de R y programadas por Marc Schwartz, permiten obtener las estadísticas γ, τ_b y τ_c

```
concordant <- function(x)
{
  x <- matrix(as.numeric(x), dim(x))
  mat.lr <- function(r, c)
  {
    lr <- x[(r.x > r) & (c.x > c)]
    sum(lr)
  }
  r.x <- row(x)
  c.x <- col(x)
 sum(x * mapply(mat.lr, r = r.x, c = c.x))
}

discordant <- function(x)
{
  x <- matrix(as.numeric(x), dim(x))
  mat.ll <- function(r, c)
  {
    ll <- x[(r.x > r) & (c.x < c)]
```

```
    sum(ll)
  }
  r.x <- row(x)
  c.x <- col(x)
  sum(x * mapply(mat.ll, r = r.x, c = c.x))
}
calc.KTb <- function(x)
{
  x <- matrix(as.numeric(x), dim(x))
  c <- concordant(x)
  d <- discordant(x)
  n <- sum(x)
  SumR <- rowSums(x)
  SumC <- colSums(x)
  Wf<-( n ^ 2 -sum( SumR^2 ) )/2
  Wc<-( n ^ 2  -sum( SumC^2) )/2
  KTb <- (c - d)/ sqrt(Wf*Wc)
  list(KTb=KTb,Wf=Wf,Wc=Wc)
}
```

Tabla 2.9

```
t2.9<-t2.1[,c(3,2,1)]
D<-discordant(t2.9)
C<-concordant(t2.9)
# gamma estimado
(C-D)/(C+D)
#Estadístico tau b
calc.KTb(t2.9)
#Estadístico Tau c
N<-sum(t2.9)
k<-min(nrow(t2.9),ncol(t2.9) )
(C-D)/(N^2*(k-1)/(2*k) )
#Estadística de somers
Wf<-calc.KTb(t2.9)$Wf
Wc<-calc.KTb(t2.9)$Wc
# D(Y|X)
(C-D)/Wf
# D(X|Y)
(C-D)/Wc
```

Coeficiente de correlación de Pearson calculado a partir de la tabla 2.2 (sección 2.4.8).

```
fi<-1:4 # puntajes fila
ci<-1:3 # puntajes columna
n<-sum(t2.9) # N
ni.<-margin.table(t2.9,1)
fbar<-sum(ni.*fi)/n # ec 2.31a
n.j<-margin.table(t2.9,2)
cbar<-sum(n.j*ci)/n # ec 2.31b
desvc<-matrix(rep(ci-cbar,4),nrow=4
              ,byrow=TRUE)
desvf<-matrix(rep(fi-fbar,3),nrow=4)
numrho<-sum(t2.9*desvc*desvf) # numerador para rho
denrho1<-sum(t2.9*matrix(rep((fi-fbar)^2,3),nrow=4))
denrho2<-sum(t2.9*matrix(rep((ci-cbar)^2,4),
             nrow=4,byrow=TRUE))
numrho/sqrt(denrho1*denrho2) # rho
```

Tablas de contingencia 2×2, prueba ji–cuadrado a partir de la tabla 2.17.

```
t2.15<-as.table(matrix(c(16,40,48,20),nrow=2))
# sin corrección
chisq.test(t2.15,correct=FALSE)
# con corrección
chisq.test(t2.15)
```

Prueba exacta de Fisher (sección 2.5.3)

```
t2.16<-as.table(matrix(c(10,2,2,4),nrow=2))
fisher.test(t2.16) # bilateral
fisher.test(t2.16,alt ="less") # unilateral inferior
fisher.test(t2.16,alt="greater") # unilateral superior
```

Prueba de McNemar a partir de la tabla 2.21

```
# tabla 2.19
t2.19<-as.table(matrix(c(14,2,5,2),nrow=2))
```

```
# sin corrección
mcnemar.test(t2.19,correct = FALSE)
# con corrección
mcnemar.test(t2.19)
# tabla 2.20
t2.20<-as.table(matrix(c(9,13,3,75),nrow=2))
mcnemar.test(t2.20)
```

Riesgo relativo (sección 2.5.5), se requiere la librería `epibasix`.

```
t2.21<-matrix(c(189,104,10845,10933),nrow=2)
require(epibasix)
ant2.21<-epi2x2(t2.21)
summary(ant2.21)
names(ant2.21)
```

Prueba de Mantel–Haenszel a partir de la tabla 2.28

```
t2.23<-array(c(29,14,16,31,
        37,24,8,21,
        30,23,15,22),
dim=c(2,2,3),
dimnames=list(Treatment=c("Medicamento","Placebo"),
Mejoria=c("Si","No"), centro=c("1","2","3")))
addmargins(t2.23)
mantelhaen.test(t2.23,correct = FALSE)  # sin corrección
```

2.9 Ejercicios

1. A partir de una muestra de 300 accidentes automovilísticos se clasificó la información de acuerdo con el tamaño del automóvil y el número de muertes ocurridas. Los datos se muestran en la tabla 2.31, con base en esta información:

   a) ¿Depende el número de muertes del tamaño del automóvil? Justifique.

   b) Obtenga las diferentes medidas de asociación y discuta los resultados.

Tabla 2.31: Datos sobre accidentes automovilísticos.

	Tamaño del auto		
No. muertos	Pequeño	Mediano	Grande
Uno o más	42	35	20
Cero	78	65	60

2. La tabla 2.32 contiene datos sobre el uso de marihuana por estudiantes de secundaria relacionados con el uso de alcohol y drogas por sus padres.

   a) ¿Es independiente el uso de marihuana por parte de los estudiantes del uso de alcohol y drogas por parte de sus padres? Justifique mediante la prueba ji–cuadrado y mediante la prueba de la razón de verosimilitudes.

   b) Obtenga las diferentes medidas de asociación y discuta los resultados.

Tabla 2.32: Datos sobre uso de marihuana por estudiantes.

	Estudiante		
Padres	Nunca	Ocasionalmente	Regularmente
Ambos	17	11	19
Ninguno	141	54	40
Uno	68	44	51

3. La tabla 2.33 contiene los resultados de un estudio para comparar los resultados del uso de la terapia de radiación frente a la cirugía en el tratamiento de cierto cáncer de laringe. Use la prueba exacta de Fisher para decidir si existe asociación entre el tratamiento y el control o no del cáncer.

Tabla 2.33: Comparación entre radiación y cirugía en el tratamiento de cáncer de laringe.

	Cáncer controlado	
Tratamiento	Sí	No
Cirugía	21	2
Radiación	15	3

4. La tabla 2.34 contiene los datos de 20 pacientes intervenidos quirúrgicamente, en los que se valoró el dolor después de la cirugía y al cabo de 1 hora tras la administración de un analgésico. Aplique la prueba de McNemar e interprete los resultados.

Tabla 2.34: Datos de dolor tras la cirugía.

	Dolor una hora después	
Dolor tras la intervención	Sí	No
Sí	1	11
No	2	6

Capítulo 3

Análisis de correspondencias

3.1 Introducción

El *análisis de correspondencias* se desarrolla mediante el trabajo sobre dos tablas de datos: la primera tabla contiene las frecuencias respecto a las modalidades de dos variables; usualmente se denomina *análisis de correspondencias binarias*. El segundo tipo de tabla contiene la información sobre diversas variables; el análisis se conoce como de *correspondencias múltiples*.

Como ejemplo, considere la matriz de frecuencias (n_{ij}) contenida en la tabla 3.1, tomada de Thompson (1995). En esta tabla, las filas $(i = 1, 2, 3, 4)$ son el color de los ojos y las columnas $(j = 1, 2, 3, 4, 5)$ el color del cabello, cuyas modalidades varían de claro a oscuro. Para encontrar la representación más adecuada de estos datos, es necesario comparar las filas y las columnas de la tabla. Tal comparación implica usar una medida de distancia apropiada. El análisis de correspondencias permite describir las proximidades existentes entre los perfiles, color del cabello (perfil fila) y color de los ojos (perfil columna), de acuerdo con la partición que se haga de los individuos, sea por filas o por columnas.

La matriz de densidades o frecuencias relativas (f_{ij}) y las densidades marginales de filas $(f_{i.})$ y columnas $(f_{.j})$ es mostrada en la tabla 3.2. Los números son dados como porcentaje y representan el $f_{ij}100\%$. Los

Tabla 3.1: Frecuencias absolutas.

Color de ojos	Color del cabello Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	Total ($n_{i.}$)
Claros (C)	688	116	584	188	4	1580
Azules (A)	326	38	241	110	3	718
Medio (M)	343	84	909	412	26	1774
Oscuros (O)	98	48	403	681	85	1315
Total ($n_{.j}$)	1455	286	2137	1391	118	5387

números a la derecha de cada fila presentan las densidades marginales, como el porcentaje $f_{i.}100\%$; la última fila representa las densidades marginales por columna $f_{.j}100\%$. En resumen, la mayoría de las personas tienen el color de los ojos medio $(32,93\%)$ y el color de cabello más común es también medio $(39,66\%)$. El origen del análisis de correspondencias se puede remontar a los trabajos de Hirschfeld (1935) y de Fisher (1940) sobre tablas de contingencia, pero el verdadero responsable de esta técnica estadística es Benzecri (1973), tal como se cita en Lebart (1985, 276). En reconocimiento a la escuela francesa, se mantienen en este texto algunos de sus términos, los cuales tienen sus respectivas nominaciones en la escuela anglosajona.

Tabla 3.2: Frecuencia relativas.

Color de ojos	Color del cabello Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	Total ($f_{i.}$)
Claros (C)	12.77	2.15	10.84	3.49	0.07	29.32
Azules (A)	6.05	0.71	4.47	2.04	0.06	13.33
Medio (M)	6.37	1.56	16.87	7.65	0.48	32.93
Oscuros (O)	1.82	0.89	7.48	12.65	1.58	24.42
Total ($f_{.j}$)	27.01	5.31	39.66	25.83	2.19	100.00

3.2 Representación geométrica de los puntos de una tabla de contingencia

En una tabla de contingencia (matriz de datos) pueden considerarse dos espacios, el espacio fila ($\mathbb{R}^p$) o el espacio columna ($\mathbb{R}^n$). Para el ejemplo anterior, el espacio *color de los ojos* ($\mathbb{R}^5$) y el espacio *color del cabello* ($\mathbb{R}^4$), respectivamente.

La matriz de datos $\mathbb{X}$ (tabla de contingencia) tiene n-filas y p-columnas; n_{ij} representa el número de individuos de la fila i y la columna j. En el ejemplo, n_{ij} es el número de individuos con el color de los ojos i y color del cabello j.

El número total de individuos por fila se nota por

$$n_{i.} = \sum_{j=1}^{p} n_{ij}, \quad \text{para } i = 1, \cdots, n. \tag{3.1}$$

El número total de individuos por columna se nota por

$$n_{.j} = \sum_{i=1}^{n} n_{ij}, \quad \text{para } j = 1, \cdots, p. \tag{3.2}$$

El número total de individuos de la tabla está dado por

$$N = \sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij} = \sum_{i=1}^{n} n_{i.} = \sum_{j=1}^{p} n_{.j}. \tag{3.3}$$

Las frecuencias relativas absolutas y marginales se notan como sigue

$$f_{ij} = \frac{n_{ij}}{N}; \quad f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i.}}{N}; \text{ y } f_{.j} = \sum_{i=1}^{n} f_{ij} = \frac{n_{.j}}{N}. \tag{3.4}$$

Con lo anterior se puede apreciar que la matriz (tabla) $\mathbb{X}$ de elementos n_{ij} se ha transformado en la matriz (tabla) de elementos f_{ij}; esta última se nota por $\mathbf{F} = (f_{ij})$.

Las frecuencias relativas condicionales de columna respecto a filas (perfiles) y fila respecto a columnas se escriben, respectivamente, como sigue:

$$f_{i|j} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} \text{ y } f_{j|i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}, \text{ para } i = 1, \ldots, n \; j = 1, \ldots, p. \tag{3.5}$$

En el espacio fila ($\mathbb{R}^p$) o *nube de puntos fila,* el i-ésimo vector (perfil fila) tiene coordenadas

$$\left(\frac{n_{i1}}{n_{i.}},\ldots,\frac{n_{ip}}{n_{i.}}\right)=\left(\frac{f_{i1}}{f_{i.}},\ldots,\frac{f_{ip}}{f_{i.}}\right)=\left(f_{1|i},\ldots,f_{p|i}\right);\ i=1,\ldots,n. \tag{3.6}$$

La nube de puntos fila (perfil fila) queda representada por $\mathbf{D}_n^{-1}\mathbf{F}$, con $\mathbf{D}_n=\text{Diag}(f_{i.})$, matriz diagonal que contiene las frecuencias marginales por fila o "pesos" $f_{i.}$. Se observa que cada punto o perfil fila está afectado por su peso $f_{i.}$.

El *centroide* o *baricentro (centro de gravedad)* de la nube de puntos fila se representa por $\boldsymbol{\mathcal{G}_f}$; sus coordenadas son las frecuencias marginales, es decir, $\mathcal{G}_f=(f_{.1},\ldots,f_{.p})$.

De manera similar, en el espacio columna ($\mathbb{R}^n$) o *nube de puntos columna*, el j-ésimo vector (perfil columna) tiene coordenadas

$$\left(\frac{n_{1j}}{n_{.j}},\cdots,\frac{n_{nj}}{n_{.j}}\right)=\left(\frac{f_{1j}}{f_{.j}},\ldots,\frac{f_{nj}}{f_{.j}}\right)=\left(f_{1|j},\ldots,f_{n|j}\right);\ j=1,\ldots,p. \tag{3.7}$$

También, el *centroide* o *baricentro* de la nube de puntos columna se representa por $\boldsymbol{\mathcal{G}_c}$. Sus coordenadas son las frecuencias marginales; es decir, $\mathcal{G}_c=(f_{1.},\ldots,f_{n.})$.

En forma gráfica se puede representar lo anterior mediante el esquema de la figura 3.1.

3.2.1 Perfiles fila y columna

Las tablas 3.3 y 3.4 contienen los perfiles fila y columna, respectivamente. Así, la tabla 3.3 muestra la distribución del color del cabello por cada uno de los colores de los ojos; recíprocamente, la tabla 3.4 suministra la distribución del color de ojos manteniendo constante el color del cabello.

La distribución de frecuencias condicionadas, del color de cabello de acuerdo con el color de los ojos de las personas estudiadas, se representa en el vector ($n_{ij}/n_{i.}=f_{j|i}$), ilustrado en la figura 3.2. Alternamente, se ilustra la distribución condicional de frecuencias del color de los ojos respecto al color del cabello ($n_{ij}/n_{.j}=f_{i|j}$) en la figura 3.3.

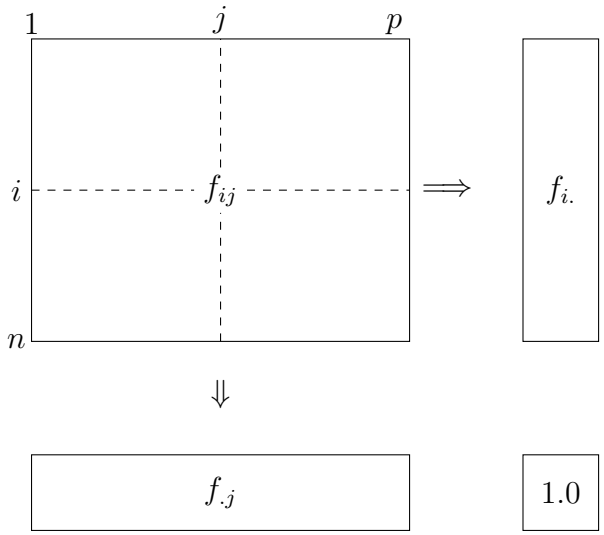

Figura 3.1: Tabla de frecuencias y sus marginales.

Tabla 3.3: Perfil fila.

Color de ojos	Color del cabello Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	Total
Claros (C)	0.4354	0.0734	0.3697	0.1190	0.0025	1.0000
Azules (A)	0.4540	0.0529	0.3357	0.1532	0.0042	1.0000
Medio (M)	0.1933	0.0474	0.5124	0.2322	0.0147	1.0000
Oscuros (O)	0.0745	0.0365	0.3065	0.5179	0.0646	1.0000
Centroide columna	0.2701	0.0531	0.3966	0.2583	0.0219	1.0000

Los perfiles fila y columna pueden ser comparados con las distribuciones columna y fila con el respectivo peso, para juzgar su "apartamiento" de la independencia. La figura del perfil *color de ojos respecto al color del cabello* muestra una alta similitud entre los perfiles ojos claros y ojos azules, igual a, aunque un poco más baja, la similitud o proximidad entre los perfiles ojos medios y oscuros (figura 3.2).

Para el perfil color del cabello, se encuentra una alta semejanza entre los perfiles cabello rubio y rojo, así como entre los cabellos oscuro y negro; el perfil cabello medio es bastante diferente a los demás, como se

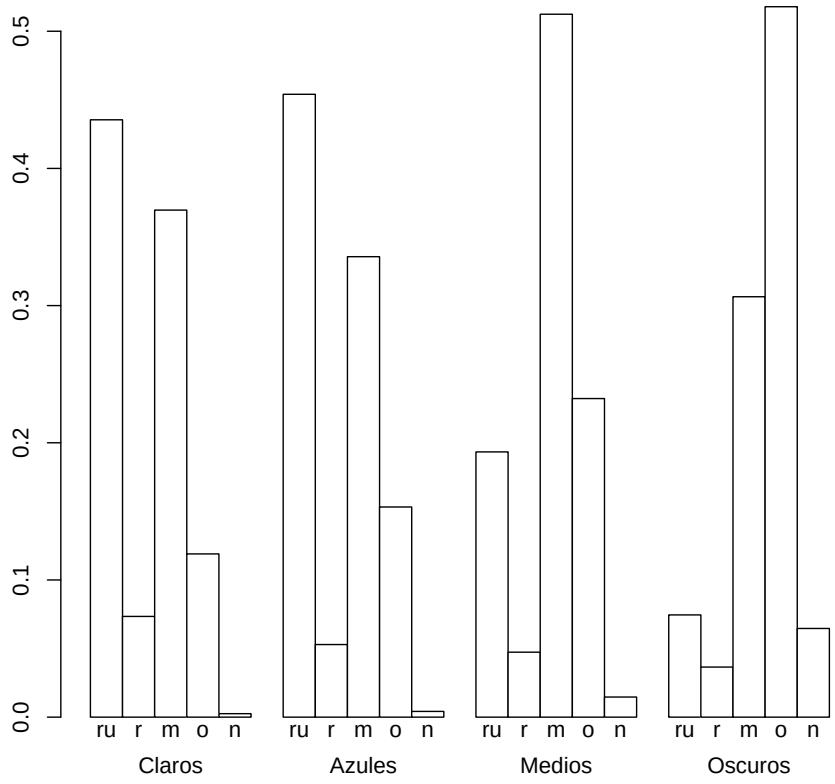

Figura 3.2: Perfiles fila.

muestra en la figura 3.3.

3.3 Semejanza entre perfiles: la distancia ji–cuadrado

Una vez que se han definido las dos nubes de puntos, espacio fila ($\mathbb{R}^p$) y espacio columna ($\mathbb{R}^n$), se decide cómo medir la distancia entre ellos. En

Tabla 3.4: Perfil columna.

Color de ojos	Color del cabello Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	Total
Claros (C)	0.4729	0.4056	0.2733	0.1352	0.0339	0.2932
Azules (A)	0.2241	0.1329	0.1128	0.0791	0.0255	0.1333
Medio (M)	0.2356	0.2937	0.4254	0.2961	0.2203	0.3293
Oscuros(O)	0.0674	0.1678	0.1885	0.4896	0.7203	0.2442
Centroide columna	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

el análisis de correspondencias, la semejanza entre dos filas (o entre dos columnas) está dada por la distancia entre sus perfiles. Esta distancia es conocida con el nombre de distancia *ji–cuadrado*; se nota χ^2. Se define en forma análoga la distancia entre perfiles fila y columna, respectivamente.

La distancia entre dos perfiles fila i e i' está dada por

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2. \tag{3.8}$$

Similarmente, la distancia entre dos perfiles columna j y j' es

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2. \tag{3.9}$$

Note que (3.8) y (3.9) miden la distancia entre dos distribuciones multinomiales; es decir, permiten comparar los histogramas (distribuciones empíricas) por cada par de filas o columnas.

Las distancias dadas en las ecuaciones (3.8) y (3.9) difieren de la distancia euclidiana en que cada cuadrado es ponderado por el inverso de la frecuencia para cada modalidad; es decir, se ponderan las distintas coordenadas, de manera que se le da más "importancia" a las categorías o modalidades con menor frecuencia y menos "importancia" a las que tengan alta frecuencia.

Las distancias anteriores se traducen en que el análisis de correspondencias da prioridad a las modalidades raras, por cuanto estas, por su

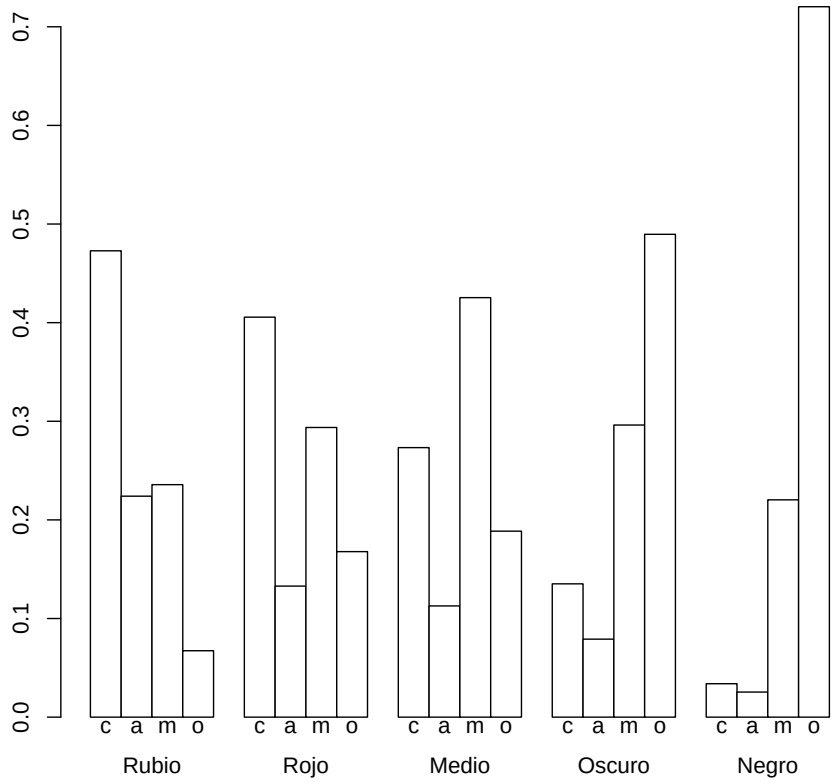

Figura 3.3: Perfiles columna.

escasez, son más diferenciadoras que las otras.

3.4 Explicación de la técnica

En lugar de comparar filas/columnas, el análisis de correspondencias procede derivando un pequeño número de dimensiones, de forma que la *primera dimensión* (*primer eje o factor*) "explique" la mayor parte de la asociación total entre las filas y las columnas (asociación medida en

términos χ^2/N), la *segunda dimensión* (*segundo eje o factor*) explique la mayor parte de la asociación no explicada por la primera (residuo), y así sucesivamente con el resto de dimensiones. El máximo número de dimensiones es igual al menor número de modalidades de cualquiera de las variables (fila o columna) menos uno, pero en general dos o tres dimensiones son suficientes para representar de modo adecuado la asociación entre ambas variables.

La cantidad χ^2/N se denomina *inercia total*. Esta inercia se descompone en un total de k valores característicos (propios), cada uno de los cuales constituye la inercia principal de una dimensión.

Cada una de las dimensiones (*ejes principales*) caracteriza las categorías de filas y columnas de la tabla de contingencia bidimensional situándolas como coordenadas en el espacio geométrico: las *puntuaciones de fila* son las coordenadas para cada modalidad de la variable fila en las dimensiones de la tabla de contingencia y las *puntuaciones columna* son las coordenadas para cada modalidad de la variable columna en esas mismas dimensiones.

Para encontrar las coordenadas de las modalidades de las variables de fila y columna se procede, de manera esquemática, como se muestra a continuación:

- La primera dimensión o eje principal produce un valor singular (valor propio) λ_1 que explica la mayor parte de la inercia (asociación) total y constituye el primer conjunto de coordenadas: uno para la variable de fila y otro para la variable de columna.
- La segunda dimensión produce un nuevo valor singular λ_2 que explica la mayor parte de la inercia residual y constituye el segundo conjunto de coordenadas: una para la variable fila y otra para la variable columna.
- El resto de dimensiones se obtiene con el mismo procedimiento.

Una explicación más detallada sobre esta metodología se puede consultar en Díaz (1999: cap. 11).

Para ilustrar la técnica del análisis de correspondencias, se toman los datos de la tabla 3.1. La tabla de contingencia para el color de ojos y cabello en una muestra de 5.387 personas, nuevamente, es la siguiente. Los valores característicos (o valores propios) son, en forma decreciente,1.0000,

Tabla 3.5: Color de ojos vs. color del cabello.

Color de ojos	Color del cabello Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	Total
Claros (C)	688	116	584	188	4	1580
Azules (A)	326	38	241	110	3	718
Medio (M)	343	84	909	412	26	1774
Oscuros (O)	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	53874

0.1992, 0.0301, 0.0009 y 0.0000. Se descarta el valor propio igual a 1.0000. En el siguiente cuadro se resumen los valores propios junto con la inercia individual y acumulada retenida por cada valor propio.

```
Valor Propio Porcentaje Porc. Acum.
 0.1992       86.56       86.56        * * * * * * * * *
 0.0301       13.07       99.63        * * *
 0.0009        0.37      100.00        *
 0.0000        0.00      100.00        *
```

Del cuadro anterior se lee que con la primera dimensión se reúne el $86,6\%$ de la varianza y que con la segunda dimensión se reúne casi toda su variabilidad, es decir, $99,6\%$.

Las coordenadas para las dos dimensiones, por fila y columna, se muestran en la tabla 3.6. La figura 3.4 representa la proyección de los puntos fila y columna (tabla 3.6) en el primer plano factorial. La primera dimensión está relacionada con el color del cabello, variando, de izquierda a derecha, desde el color oscuro al claro, respectivamente. Se puede apreciar que los datos referentes a los ojos siguen un "patrón" similar al del cabello, con colores oscuros a la izquierda y claros a la derecha. Los puntos para *azul* y *rubio* están razonablemente próximos; aunque algunas veces es difícil determinar si las personas tienen ojos claros o azules por problemas de pigmentación. En resumen, la dirección del color va de izquierda a derecha, y de *claro* a *oscuro*, tanto para el cabello como para los ojos. El procedimiento para el análisis de correspondencias simple o binaria se puede resumir en las siguientes etapas, las cuales se ilustran

Tabla 3.6: Coordenadas, color de ojos y del cabello.

Coordenadas fila Color de ojos			Coordenadas columna Color del cabello		
	Dim.1	Dim.2		Dim.1	Dim.2
Claros	0.44	0.09	Rubio	0.54	0.17
Azules	0.40	0.17	Rojo	0.23	0.05
Medios	−0.03	−0.24	Medio	0.04	−0.21
Oscuros	−0.70	0.13	Oscuro	−0.59	0.10
			Negro	−1.09	0.29

en la figura 3.5.

1. Se parte de los datos originales; las filas y columnas cumplen papeles simétricos. Estas son las modalidades de las dos variables, respectivamente. La suma de todos los términos de la tabla es n, el cual es el número total de individuos o efectivos.

2. Se construye una tabla de las frecuencias relativas las cuales conforman las probabilidades. Donde $(f_{i.} : \ i = 1, \ldots, n)$ y $(f_{.j} : j = 1, \ldots, p)$ son las probabilidades marginales o perfiles fila o columna, respectivamente.

3. Para estudiar las líneas de la tabla, se les transforma en perfiles fila. Análogamente se procede con las columnas. Se dispone entonces de dos tablas: una para el perfil fila y otra para el perfil columna. Un perfil se interpreta como una probabilidad condicional. El perfil medio es la distribución asociada a la que se presenta en el numeral 2.

4. Un perfil-fila es un arreglo de p-números y está representado por un punto de $\mathbb{R}^p$. La nube de puntos $\mathcal{H}_c$, de los perfiles fila, está en un hiperplano $\mathcal{H}_f$ de vectores tales que la suma de sus componentes es igual a 1. Cada perfil fila i es afectado por los puntos $f_{i.}$; de manera que la nube $\mathcal{H}_f$ está "equilibrada" en los perfiles medios o baricentro $\mathcal{G}_i$. En la nube $\mathcal{H}_f$ se busca la semejanza entre los perfiles, semejanza medida a través de una distancia χ^2.

5. La representación de los perfiles columna de $\mathbb{R}^n$ se hace de forma análoga a la representación de los perfiles fila en $\mathbb{R}^p$.

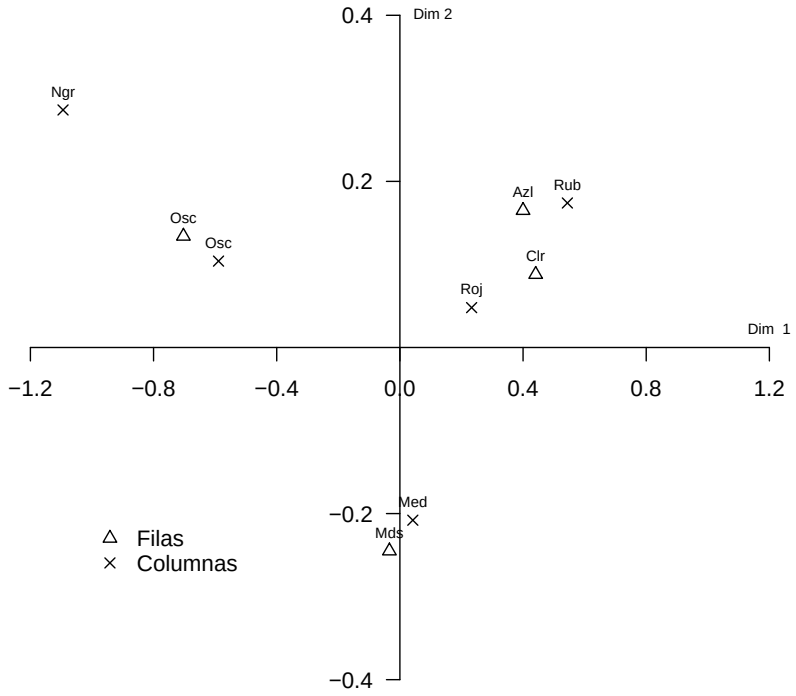

Figura 3.4: Proyección: datos color de ojos (△) y del cabello (×).

6. El análisis factorial de la nube consiste en poner en evidencia una sucesión de direcciones ortogonales, tales que la inercia, con relación al origen O de la proyección de la nube de puntos sobre tales direcciones, sea máxima.

7. Simétricamente, se desarrolla un procedimiento análogo para las columnas.

8. Los planos factoriales, determinados por dos factores sobre las filas o sobre las columnas, proporcionan imágenes aproximadas de las

nubes $\mathcal{H}_f$ y $\mathcal{H}_c$; sobre este plano, la distancia entre dos puntos se interpreta como la semejanza entre los perfiles de esos puntos. El origen de los ejes se considera como el perfil promedio.

9. Las relaciones de transición expresan los resultados de un análisis factorial, por ejemplo los del espacio fila en función del espacio columna, y recíprocamente, los del espacio columna en función del espacio fila.

10. Una vez realizadas las transiciones, las interpretaciones de los planos factoriales que representan a $\mathcal{H}_f$ y $\mathcal{H}_c$ deben hacerse conjuntamente. Esta es la comodidad de las superposiciones, la interpretación de esta representación simultánea se facilita por la propiedad del doble baricentro.

3.5 Análisis de correspondencias múltiples

El AC se ha ocupado, principalmente, de tablas de contingencia bidimensionales. El análisis de correspondencias puede extenderse a tablas de tres o más entradas. Las filas de estas tablas son los objetos o individuos y las columnas las modalidades de variables categóricas. Es el caso de las encuestas, donde las filas son individuos, grupos humanos o instituciones; y las columnas, modalidades de respuesta a las preguntas formuladas en el cuestionario o instrumento. El *análisis de correspondencias múltiples* es un análisis de correspondencias simple aplicado no solo a una tabla de contingencia, sino a una tabla disyuntiva completa.

3.5.1 Tablas de datos

A manera de ilustración, considere un conjunto de n individuos a los cuales se les registra:

El grupo de edad. Modalidades: joven (1), adulto (2), anciano (3)

Género. Modalidades: masculino (1), femenino (2)

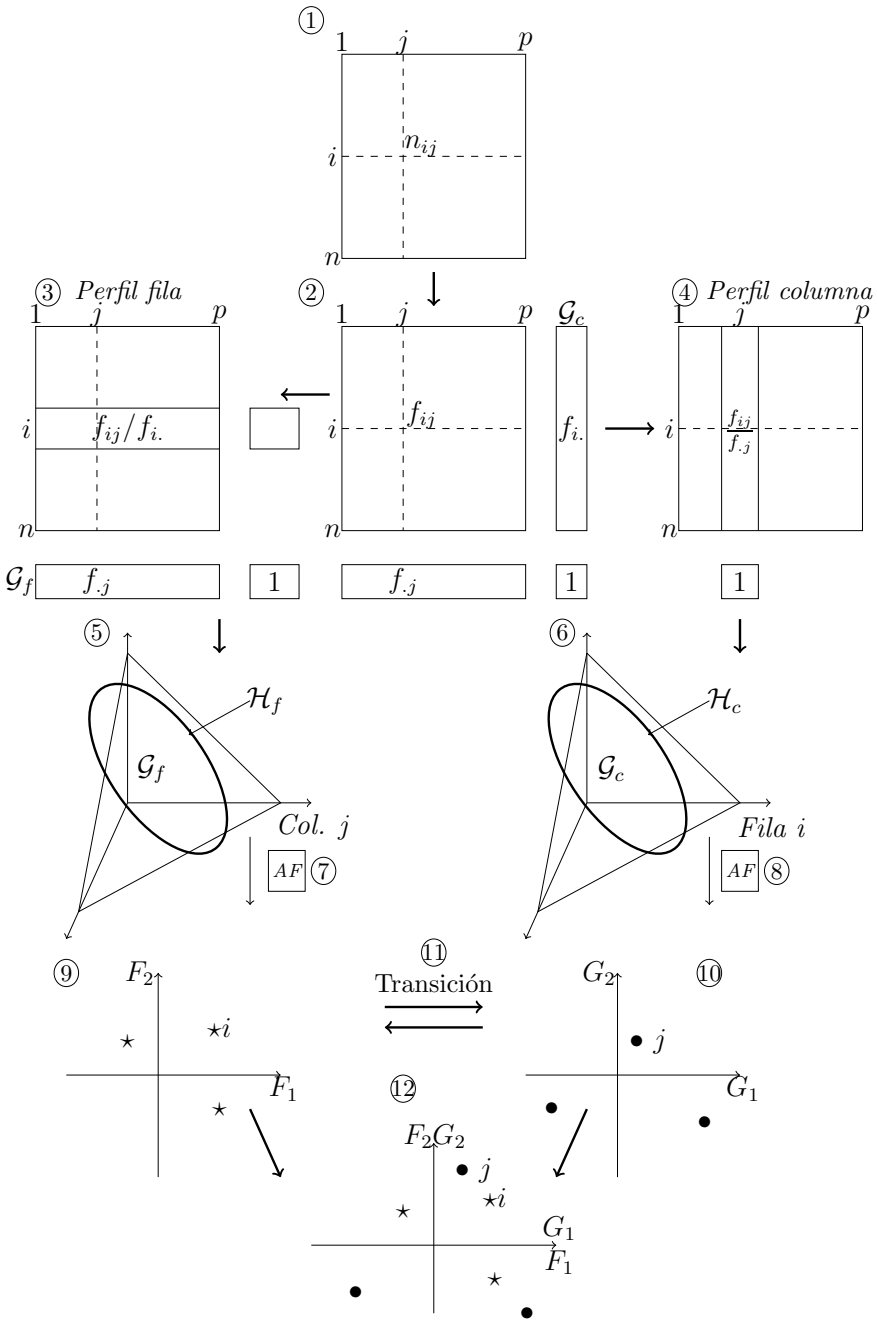

Figura 3.5: Esquema del análisis de correspondencias (tomado de Escofier y Pagés, 1990, pág. 42).

Nivel de estudios. Modalidades: primaria (1), secundaria (2), universitaria (3), otra (4)

Categoría socioeconómica. Modalidades: bajo (1), medio, (2), alto (3)

Posesión de vivienda. Modalidades: propietario (1), no propietario (2).

Se tiene entonces una matriz de datos $\mathbf{R}$ con n filas (individuos) y cinco columnas. Las entradas de esta matriz son los códigos asociados a cada modalidad de respuesta por pregunta. La siguiente es una de las matrices que surge de las posibles modalidades asumidas por los n individuos:

$$\mathbf{R} = \begin{bmatrix} 2 & 1 & 2 & 2 & 1 \\ 1 & 2 & 3 & 3 & 2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 3 & 1 & 4 & 3 & 1 \end{bmatrix}$$

Así, la primera fila de la matriz $\mathbf{R}$ señala a un hombre adulto, con estudios de secundaria, de estrato socioeconómico medio quien tiene vivienda propia.

Esta matriz o tabla de datos no es tratable vía análisis de correspondencias múltiples; pues la suma de estos números en filas o en columnas no tienen ningún sentido. Una salida para el análisis de esta tabla es una recodificación. Esta recodificación se logra cruzando los individuos con las combinaciones de modalidades para cada una de las preguntas; para el caso se tienen 5 preguntas con 3, 2, 4, 3 y 2 modalidades, respectivamente, es decir, $3 \cdot 2 \cdot 4 \cdot 3 \cdot 2 = 144$ posibles respuestas de los individuos.

Mediante el uso de variables indicadoras se convierte una tabla múltiple en una tabla de doble entrada. Suponga, en general, una tabla con k-variables (o preguntas) donde cada una de las cuales tiene p_i *modalidades* o *categorías* (para $i = 1, \ldots, k$). Se asocia una variable indicadora por modalidad dentro de cada variable o entrada de la tabla. La codificación dada por p_i hace corresponder tantas variables binarias como modalidades tenga la variable categórica. El total de modalidades es igual a $\sum_{i=1}^{k} p_i = p$.

Para un individuo particular se codifica con uno (1) si el individuo posee el atributo de la respectiva modalidad y con cero (0) en las demás modalidades de la misma variable, pues se asume que las modalidades son excluyentes. Resulta entonces una matriz $\mathbb{X}$ de tamaño $(n \times p)$ formada por bloques columna, cada uno de los cuales hace referencia a una variable registrada sobre los n individuos.

Para la matriz $\mathbf{R}$ anterior, la codificación se muestra en la figura 3.6.

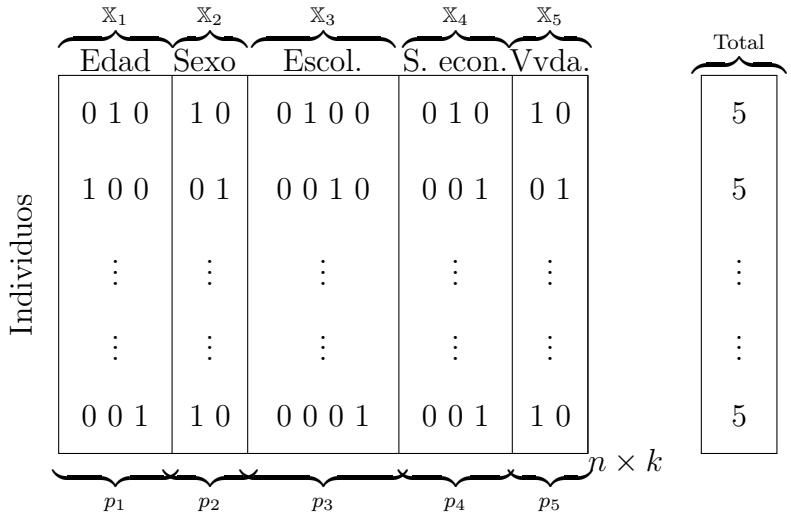

Figura 3.6: Tabla múltiple.

La suma en cada una de las filas es constante e igual a $p = 5$, mientras que la suma en las columnas n_j $(j = 1, \ldots, 14)$ suministra el número de individuos que participan en cada una de las 14 modalidades. La tabla o matriz $\mathbb{X}$ con n-filas y p-columnas describe las k-respuestas para los n-individuos a través de un código binario (0 o 1) y se le llama *tabla disyuntiva completa*. Esta tabla es la unión de k tablas (una por pregunta). Así, para el ejemplo anterior $\mathbb{X} = [\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \mathbb{X}_4, \mathbb{X}_5]$. En general,

$$\mathbb{X} = [\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_k]. \tag{3.10}$$

Cada una de las tablas $\mathbb{X}_j$, $j = 1, \ldots, k$, describe la partición de los n individuos de acuerdo con sus respuestas a la pregunta j. De otra

manera $\mathbb{X}_j = (x_{im})$, donde

$$x_{im} = \begin{cases} 1, & \text{si el } i\text{-ésimo individuo está en la modalidad } m \text{ de} \\ & \text{la pregunta } j \\ 0, & \text{si el } i\text{-ésimo individuo no está en la modalidad } m \text{ de} \\ & \text{la pregunta } j \end{cases}$$

Tabla de Burt

Para cada pregunta o variable, sus p_j respuestas o modalidades permiten particionar la muestra en máximo p_j clases. Por ejemplo, para dos preguntas se pueden hacer dos particiones del conjunto de individuos, con lo cual se obtiene una tabla de contingencia. El análisis se puede generalizar a k particiones, donde $k \geq 2$.

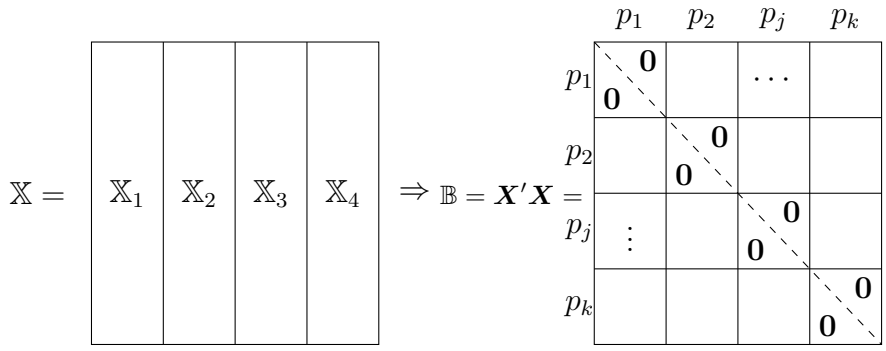

Figura 3.7: Construcción de la tabla de Burt.

A partir de la tabla disyuntiva completa $\mathbb{X}$, se construye una tabla simétrica $\mathbb{B}$ de tamaño $(p \times p)$ que contiene las frecuencias para los cruces entre todas las k variables. Esta tabla es

$$\mathbb{B} = \mathbb{X}'\mathbb{X}, \tag{3.11}$$

la cual se le conoce como *tabla de contingencia Burt* asociada a la tabla disyuntiva completa $\mathbb{X}$. Un esquema de la tabla de Burt se presenta en la figura 3.7. El término general de $\mathbb{B}$ se escribe

$$b_{jj'} = \sum_{i=1}^{n} x_{ij} x_{ij'}$$

Las marginales son

$$b_j = \sum_{j'=1}^{p} b_{jj'} = kx_{.j}, \text{ para todo } j \leq p$$

La frecuencia total es igual a

$$b = k^2 x_{.j}$$

La tabla $\mathbb{B}$ está conformada por k^2 bloques, donde:

- El bloque $\boldsymbol{X}'_j \boldsymbol{X}_{j'}$ de tamaño $(p_j \times p_{j'})$ corresponde a la tabla de contingencia que cruza las respuestas a las preguntas (variables) j y j'.
- El j-ésimo bloque cuadrado $\boldsymbol{X}'_j \boldsymbol{X}_j$ se obtiene mediante el cruce de cada variable consigo misma. Esta es una matriz diagonal de tamaño $(p_j \times p_j)$, la matriz es diagonal, dado que dos o más modalidades de una misma pregunta no pueden ser seleccionados simultáneamente. Los términos sobre la diagonal son las frecuencias de las modalidades de la pregunta j.

Sobre la diagonal de la tabla de Burt $\mathbb{B}$, en la figura 3.7 se han insinuado matrices diagonales. Estas se notan por $\mathbf{D}_j = P\mathbb{X}'_j P\mathbb{X}_j$; $j = 1, \ldots, k$ y son matrices de tamaño $(p_j \times p_j)$. Dichas matrices son diagonales puesto que un individuo no puede tener simultáneamente dos o más modalidades en una misma pregunta o variable. Los términos de la diagonal son las frecuencias de las modalidades de la pregunta j; es decir, es el número de individuos por modalidad en la pregunta j. Las matrices fuera de la diagonal de $\mathbb{B}$ son las tablas de contingencia entre las respectivas variables.

Se nota por $\mathbf{D}$ a la matriz diagonal de tamaño $(p \times p)$; es decir, sobre la diagonal están las frecuencias correspondientes a cada una de las modalidades

$$d_{jj} = b_{jj} = x_{.j}$$
$$d_{jj'} = 0 \text{ para todo } j \neq j'$$

La matriz $\boldsymbol{D}$ se puede considerar que está conformada por k^2 bloques. Las únicas matrices no nulas son las matrices diagonales $\boldsymbol{D}_j = \boldsymbol{X}'_j \boldsymbol{X}_j$; $j = 1, \ldots, k$ las cuales están dispuestas sobre la diagonal principal de $\boldsymbol{D}$.

Suponga que a un grupo de 20 individuos se le encuestó acerca de las cinco variables socioeconómicas descritas anteriormente. A continuación se muestra la matriz de datos con su código condensado $\boldsymbol{R}$, la tabla de datos disyunta completa $\mathbb{X}$, la tabla de Burt $\mathbb{B}$ y la tabla diagonal $\boldsymbol{D}$.

$$\mathbf{R} = \begin{array}{|ccccc|}\hline 2&1&2&2&1\\ 3&2&1&2&1\\ 3&1&4&2&1\\ 3&2&2&2&1\\ 2&1&1&2&1\\ 1&1&1&1&2\\ 2&1&2&2&2\\ 2&2&2&2&1\\ 3&1&3&3&1\\ 2&2&4&2&1\\ 1&2&3&3&1\\ 1&1&2&3&2\\ 2&2&2&2&1\\ 3&2&2&2&1\\ 3&1&4&3&1\\ 1&2&2&2&1\\ 3&2&3&2&1\\ 3&1&1&1&2\\ 2&2&2&1&2\\ 1&1&3&2&1\\ \hline \end{array} \quad \mathbb{X} = \begin{array}{|ccc|cc|cccc|ccc|cc|}\hline 0&1&0&1&0&0&1&0&0&1&0&1&0\\ 0&0&1&0&1&1&0&0&0&0&1&0&1&0\\ 0&0&1&1&0&0&0&0&1&0&1&0&1&0\\ 0&0&1&0&1&0&1&0&0&0&1&0&1&0\\ 0&1&0&1&0&1&0&0&0&0&1&0&1&0\\ 1&0&0&1&0&1&0&0&0&1&0&0&0&1\\ 0&1&0&1&0&0&1&0&0&0&1&0&1&0\\ 0&1&0&0&1&0&1&0&0&0&1&0&1&0\\ 0&0&1&1&0&0&0&1&0&0&1&0&1&0\\ 0&1&0&0&1&0&0&0&1&0&1&0&1&0\\ 1&0&0&0&1&0&0&1&0&0&1&0&1&0\\ 1&0&0&1&0&0&1&0&0&0&1&0&0&1\\ 0&1&0&0&1&0&1&0&0&0&1&0&1&0\\ 0&0&1&0&1&0&1&0&0&0&1&0&1&0\\ 0&0&1&1&0&0&0&0&1&0&1&0&1&0\\ 1&0&0&0&1&0&1&0&0&0&1&0&1&0\\ 0&0&1&0&1&0&0&1&0&0&1&0&1&0\\ 0&0&1&1&0&1&0&0&0&1&0&0&0&1\\ 0&1&0&0&1&0&1&0&0&1&0&0&0&1\\ 1&0&0&1&0&0&0&1&0&0&1&0&1&0\\ \hline \end{array}$$

La tabla de Burt $\mathbb{B}$ y la matriz diagonal $\boldsymbol{D}$ son, respectivamente,

$$\mathbb{B} = \begin{array}{|ccc|cc|cccc|ccc|cc|}\hline 5&0&0&3&2&1&2&2&0&1&2&2&3&2\\ 0&7&0&3&4&1&5&0&1&1&6&0&5&2\\ 0&0&8&4&4&2&2&2&2&1&5&2&7&1\\ \cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots\\ 3&3&4&10&0&3&3&2&2&2&5&3&6&4\\ 2&4&4&0&10&1&6&2&1&1&8&1&9&1\\ \cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots\\ 1&1&2&3&1&4&0&0&0&2&2&0&2&2\\ 2&5&2&3&6&0&9&0&0&1&7&1&6&3\\ 2&0&2&2&2&0&0&4&0&0&2&2&4&0\\ 0&1&2&2&1&0&0&0&3&0&2&1&3&0\\ \cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots\\ 1&1&1&2&1&2&1&0&0&3&0&0&0&3\\ 2&6&5&5&8&2&7&2&2&0&13&0&12&1\\ 2&0&2&3&1&0&1&2&1&0&0&4&3&1\\ \cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots\\ 3&5&7&6&9&2&6&4&3&0&12&3&15&0\\ 2&2&1&4&1&2&3&0&0&3&1&1&0&5\\ \hline \end{array}$$

$$\Downarrow$$

$$\mathbf{D} = \left[\begin{array}{ccc|cc|cccc|ccc|c}
5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
0 & 0 & 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 9 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 15 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5
\end{array}\right]$$

3.5.2 Fundamentos del análisis de correspondencias múltiples

Como en el análisis de correspondencias binarias, el análisis para el caso múltiple considera la nube de puntos fila (n puntos) y la nube de puntos columna (p puntos). En el primer caso, los ejes son las variables; en el segundo, los individuos. En consecuencia, los principios de ACM son iguales los del AC simples. Estos son (Lebart 1985, 113–127):

- Transformar la tabla de datos en perfiles fila y perfiles columna.
- Ajustar los datos o puntos ponderados por sus perfiles marginales (fila o columna).
- La distancia entre perfiles está dada por la ji–cuadrado.

A pesar de sus semejanzas con el análisis de correspondencias simple, el ACM tiene algunas particularidades, debido a la naturaleza misma de la tabla disyuntiva completa ($\mathbb{X}$). Se enuncian los principios del ACM

cuando este se desarrolla sobre la tabla disyuntiva completa y después se muestra la equivalencia con el análisis de la tabla de Burt ($\mathbb{B}$).

La nube de modalidades en $\mathbb{R}^n$ se puede descomponer en *subnubes*; así, la j-ésima nube corresponde al conjunto de las p_j modalidades de la variable j. Estas subnubes tienen su centro de gravedad en $\mathcal{G}_f$, el mismo de la nube global.

En resumen, el análisis de correspondencias múltiples se dirige a buscar aquellas variables o factores "cercanas" (altamente correlacionadas) a todos los grupos de modalidades. El factor F_1 representa el primer factor común al conjunto de variables categóricas iniciales. Los demás factores se obtienen con la condición de ortogonalidad sobre los anteriores.

Los factores $F_1,\ F_2, \ldots, F_k$, ubicados en el espacio de las modalidades, son los ejes en el espacio de los individuos; de tal forma que su proyección sobre estos "nuevos" ejes retienen la máxima variabilidad. Se puede observar la similitud, por lo menos conceptual, con el análisis de componentes principales, con una importante diferencia, y es que aquí cada variable está constituida por un subgrupo de variables binarias.

3.5.3 Propiedades del análisis de correspondencias múltiples

1. Es una representación gráfica de la asociación entre variables categóricas dos a dos; en consecuencia, el análisis de correspondencias simple es un caso especial para un par de variables en particular.

2. A diferencia del análisis de componentes principales, los primeros ejes, aún en forma creciente, explican una pequeña parte de la variabilidad total.

3. La distancia de una modalidad al origen en el ACM es inversamente proporcional a su participación n_j. Es decir, modalidades con participación baja (n_j pequeño) aparecen más alejadas del origen que las modalidades de mayor frecuencia.

4. Las modalidades o categorías de una variable están centradas; es decir, el centro de las modalidades de una misma variable es el origen del "nuevo" sistema de coordenadas. Así, las modalidades de una variable dicotómica se ubicarán en forma opuesta al origen.

5. El ACM es una descomposición de la nube de puntos de la varianza o inercia total del espacio de individuos (filas) o del espacio de las modalidades (columnas), en ciertas direcciones ortogonales, de tal forma que en cada dirección se maximice la inercia explicada.

6. Así como en el ACP la influencia de cada variable está dada por su varianza, las modalidades situadas a mayor distancia tienen la mayor inercia, luego son las más influyentes y, de acuerdo con la propiedad (3.), son las que tienen menor número de individuos.

7. Tal como en el AC simple, existe una relación de transición entre la "nueva" variable del espacio de los individuos y la de las modalidades.

8. La proyección de un individuo es el centro de gravedad de las modalidades que este ha escogido (a una distancia $\frac{1}{\sqrt{\lambda_\alpha}}$ del origen). Simétricamente, la proyección de una modalidad es el centro de gravedad de los individuos que la han escogido (a una distancia $\frac{1}{\sqrt{\lambda_\alpha}}$ del origen).

3.5.4 Reglas de interpretación

Decir que existen afinidades entre respuestas, equivale a decir que hay individuos que han seleccionado simultáneamente todas o casi todas, las mismas respuestas.

El análisis de correspondencias múltiples pone en evidencia a los individuos con perfiles semejantes respecto a los atributos seleccionados para su descripción. De acuerdo con las distancias entre elementos de la tabla disyuntiva completa y las relaciones baricéntricas, se expresa:

- *La cercanía entre individuos en términos de semejanzas*; es decir, dos individuos son semejantes si han seleccionado globalmente las mismas modalidades.

- *La proximidad entre modalidades de variables diferentes en términos de asociación*; es decir, estas modalidades corresponden a puntos medios de los individuos que las han seleccionado, y son próximas porque están ligadas a los mismos individuos o individuos parecidos.

- *La proximidad entre dos modalidades de una misma variable en términos de semejanza*; por construcción, las modalidades de una misma variable son excluyentes. Si ellas están cerca, su proximidad se interpreta en términos de semejanza entre los grupos de individuos que las han seleccionado (con respecto a las otras variables activas del análisis).

Las reglas de interpretación de los resultados, como coordenadas, contribuciones, cosenos cuadrados, son casi los mismos que los dispuestos para el análisis de correspondencias simples.

La noción de variable debe ser tenida en cuenta al momento de la interpretación; esta se debe hacer a través de las modalidades que la conforman. La contribución de una variable a un factor α se calcula sumando las contribuciones de las respectivas modalidades sobre ese factor. Así, se debe prestar atención a las variables que participan en la definición del factor, de acuerdo con las modalidades más "responsables" de los ejes factoriales.

Individuos y variables suplementarios

La utilización de elementos suplementarios en el ACM, sean individuos o variables , permite considerar información adicional que facilita la búsqueda de una tipología de los elementos activos; siempre que se conozcan las características de los individuos (o variables) suplementarios. Los elementos suplementarios se hacen intervenir en una tabla disyuntiva completa para:

- Enriquecer la interpretación de los ejes mediante variables que no han participado en su conformación. Se proyectará entonces en el espacio de las variables los centros de grupos de individuos definidos por las modalidades de variables suplementarias.
- Adoptar una óptica de pronóstico, proyectando las variables suplementarias en el espacio de los individuos; las variables activas representan el papel de variables explicativas. Se pueden proyectar a los individuos suplementarios en el espacio de las variables, para ubicarlos con respecto a los individuos activos o con respecto a grupos de individuos activos a manera de discriminación o separación de grupos.

Se consideran los datos sobre los 20 individuos a quienes se les registraron las variables: grupo etáreo, género, escolaridad, estrato socioeconómico y posesión de vivienda. El análisis se hace a través del procedimiento CORRESP del paquete SAS. Se construyen algunas tablas de contingencia y se determinan los factores, que junto con algunos indicadores sirven para interpretar y juzgar la calidad de los ejes factoriales. A pesar de insistir en la idealización o simulación de los datos, se aventuran algunas conclusiones derivadas del análisis de correspondencias múltiple para estos datos.

```
Valor    Porcent.   Porcent.
propio              Acumul.     5    10   15   20   25
0.64761  23.97%   23.97%    ************************
0.62382  22.24%   46.21%    **********************
0.57554  18.93%   65.14%    ******************
0.47050  12.65%   77.79%    *************
0.39452   8.89%   86.68%    *********
0.37784   8.16%   94.84%    ********
0.30070   5.16%  100.00%    *****
```

En la tabla anterior se observan siete valores propios no nulos, pues el número de variables activas es $k = 4$ y el número de modalidades es $p = 3 + 2 + 4 + 2 = 11$, de donde $p - s = 7$. Aunque no se consignó aquí, la inercia ligada a cada valor propio varía entre 0.41940 para el valor propio más grande y 0.09042 para el más pequeño. Esto no debe sorprender, ya que los códigos binarios asignados a las modalidades de una misma variable resultan, así sea artificialmente, ortogonales. Ya se advirtió sobre el cuidado de emplear los valores propios y las tasas de inercia como indicadores del número de ejes apropiados; sin embargo, a pesar de los casi siempre resultados pesimistas encontrados con estos, pues observe que con los dos primeros ejes reúnen el $46,21\%$ de la inercia total. Para efectos de interpretación de los datos, se puede y se debe hacer el análisis sobre el primer plano factorial y sobre otros planos, como el *factor 1* vs el *factor 3*, por ejemplo.

La tabla 3.7 contiene las variables, las modalidades con sus respectivas etiquetas, las coordenadas de las modalidades sobre los dos primeros factores y los cuadrados de los cosenos de las modalidades sobre los dos primeros ejes factoriales.

Tabla 3.7: Coordenadas y contribuciones de las modalidades.

Variable	Modalidad	Factor 1	Factor 2	Cosenos cuadrados	
Edad					
	∘ joven	0.57924	0.49117	0.111839	0.080417
	⊙ adulto	0.19298	-1.01751	0.020054	0.557487
	⊛ viejo	-0.53088	0.58334	0.187891	0.226857
Género					
	♠ hombre	0.51883	0.53084	0.269182	0.281789
	♡ mujer	-0.51883	-0.53084	0.269182	0.281789
Escolaridad					
	⊡ prima.	1.01657	0.72719	0.258352	0.132200
	⊟ secun.	0.16571	-0.91692	0.022466	0.687880
	⊞ univer.	-0.70487	1.00251	0.124211	0.251254
	⊠ otro	-0.91271	0.44450	0.147006	0.034868
Vivienda					
	$\widehat{\Pi}$ propie.	-0.50180	0.02075	0.755407	0.001291
	Π noprop.	1.50540	-0.06224	0.755407	0.001291
Variable suplementaria					
Estrato SE.					
	⊖ bajo	1.48530	0.20099	0.389312	0.007129
	⊘ medio	-0.29588	-0.28078	0.162585	0.146414
	⊕ alto	-0.15236	0.76180	0.005803	0.145085

Con relación al primer factor, se nota que está definido por la posesión de vivienda. Esta situación se corrobora con los cosenos cuadrados; recuerde que un valor de estos cercano a 1.0 indica un ángulo de la modalidad con el respectivo eje próximo a 0.0, es decir, una alta asociación entre la modalidad y el eje. También se destaca la diferenciación mostrada entre el grupo etáreo "viejo" y los demás, con una proximidad a la posesión de vivienda, lo que sugiere una relación directa entre la tenencia de vivienda y la edad. Una conclusión similar se puede establecer para la edad y el nivel de escolaridad: los datos exhiben que el nivel de escolaridad superior (universitaria y otro) están asociadas a edades avanzadas.

El segundo factor está determinado por la escolaridad superior y secundaria. La variable suplementaria, nivel socioeconómico, refuerza la asociación de este eje con tales aspectos. Respecto al género se puede afirmar, desde estos datos, que no definen los ejes (se ubican en la bisectriz

principal). Para la variable edad, la modalidad "joven" es indiferente en la definición de alguno de los dos ejes (se ubica en la bisectriz principal); en cambio las modalidades adulto y viejo son opuestas y están altamente ligadas con el segundo eje. La figura 3.8 muestra la disposición de las modalidades en el primer plano factorial. Se observa que el primer eje factorial (factor 1) está altamente determinado por la variable posesión de casa propia. Así, este eje determina dos tipologías de individuos: del *lado izquierdo* están quienes poseen un nivel de escolaridad universitario

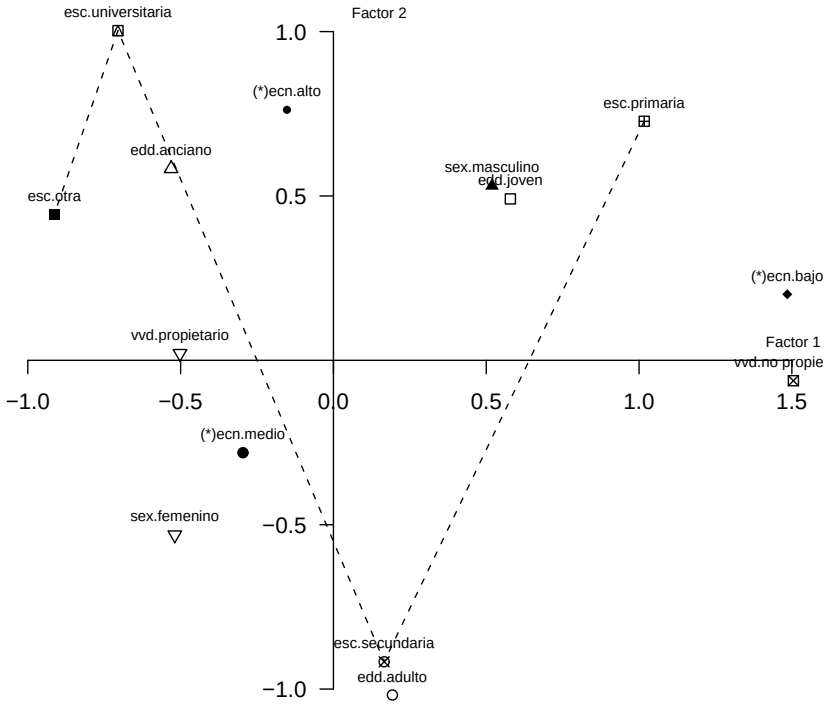

Figura 3.8: Variables activas y suplementarias en el primer plano factorial.

o más, con vivienda propia y, por tanto, en estratos socioeconómicos medios y altos, mientras que del *lado derecho* se encuentran quienes tienen un nivel de escolaridad a lo más de secundaria y que no tienen vivienda propia. El segundo eje factorial (factor 2) está definido por las modalidades asociadas a la escolaridad, discriminada esta por la modalidad secundaria frente a la modalidad universitario u otro. Se han unido, en forma ordenada, las modalidades mediante una línea poligonal. Con estas líneas se puede leer y descubrir relaciones entre las modalidades. Note, por ejemplo, que la línea que une las modalidades de escolaridad tiene un orden decreciente de izquierda a derecha (por magnitud de escolaridad) y pasa cerca o paralela a las modalidades con las que se asocia directamente. Un ejercicio similar puede hacerse con las demás modalidades.

3.6 Procesamiento de datos con R

3.6.1 Análisis de correspondencias simple

El análisis de correspondencias simple se realiza con la librería `ca`, la cual debe estar instalada previamente. Otra función para análisis de correspondencias simples es `corresp` de la librería `MASS`; para detalles de esta función consulte la ayuda mediante los comandos:
`library(MASS)`
`?corresp`.

Antes de realizar el análisis de correspondencias, se introducen los datos, se calculan las frecuencias y se realizan los gráficos de perfiles.

Introducción de la tabla 2.1 como una matriz.

```
t3.1<-matrix(c(688,116,584,188, 4,
               326, 38,241,110, 3,
               343, 84,909,412,26,
                98, 48,403,681,85),nrow=4,byrow=T)
```

Nombres de las filas y las columnas de la tabla.

```
dimnames(t3.1)<-list(col.ojos=c("Claros","Azules",
"Medios","Oscuros"),col.cabello=c("Rubio","Rojo",
"Medio","Oscuro","Negro"))
```

Se convierte la matriz en un objeto de clase tabla (table) y se agregan las marginales.

```
t3.1<-as.table(t3.1)
tabla3.1<-addmargins(t3.1)
```

Frecuencias de las celdas f_{ij}, tabla 3.2.

```
round(prop.table(t3.1)*100,2)
round(addmargins(prop.table(t3.1)*100),2)
```

Perfiles fila (tabla 3.3).

```
round(t3.3<-prop.table(t3.1,1),4)
round( addmargins(prop.table(t3.1,1)),4 )
```

Gráfico de los perfiles fila (figura 3.2).

```
barplot(t(t3.3) ,beside=T, legend.text =T )
```

Perfil columna (tabla 3.4).

```
round(t3.4<-prop.table(t3.1,2),4)
round( addmargins(prop.table(t3.1,1)),4 )
```

Perfiles columna (figura 3.3).

```
barplot(t3.4 ,beside=T,legend.text =T)
```

Análisis de correspondencias, usando la librería `ca`.

```
# se carga la librería ca
library(ca)
acs<-ca(t3.1)
summary(acs)
```

En `Rows:` de la salida proporcionada por `summary(acs)`, marcadas con `k=1` y `k=2` están las *coordenadas fila*, que se muestran en la tabla 3.6, pero multiplicadas por mil y con el signo contrario. Si se quiere recuperar esa información, como aparece en dicha tabla, se hace lo siguiente:

```
res<-summary(acs)
cord.filas<--cbind(res$rows[,5],res$rows[,8])/1000
cord.filas
```

En `Columns:` de la salida de `summary(acs)`, marcadas con `k=1` y `k=2` están las *coordenadas columna*, que se muestran en la tabla 3.6, pero multiplicadas por mil y con el signo contrario. Si se quiere recuperar esa información, como aparece en dicha tabla, se hace lo siguiente:

```
cord.col<--cbind(res$columns[,5],res$columns[,8])/1000
cord.col
```

Con la función `plot()` sobre un objeto de clase `ca` se obtiene el biplot (figura 3.4).

```
# biplot
plot(acs)
```

3.6.2 Análisis de correspondencias múltiples

El análisis de correspondencias múltiples se hace con la función `mjca()`[1] de la librería `ca`. Las librerías `MASS` y `ade4` también proporcionan funciones para análisis de correspondencias múltiples; el lector puede consultar la ayuda para mayores detalles.

Antes de realizar el análisis se introducen los datos y se definen los factores.

Introducción de la matriz $\boldsymbol{R}$.

```
edad<-c(2,3,3,3,2,1,2,2,3,2,1,1,2,3,3,1,3,3,2,1)
sexo<-c(1,2,1,2,1,1,1,2,1,2,2,1,2,2,1,2,2,1,2,1)
esco<-c(2,1,4,2,1,1,2,2,3,4,3,2,2,2,4,2,3,1,2,3)
econ<-c(2,2,2,2,2,1,2,2,3,2,3,3,2,2,3,2,2,1,1,2)
vvda<-c(1,1,1,1,1,2,2,1,1,1,1,2,1,1,1,1,1,2,2,1)
```

A continuación se convierten los vectores anteriores en factores y se organizan en un marco de datos (*data frame*).

[1] Multiple and joint correspondence analysis.

```
edad<-factor(edad,labels=c("joven","adulto","anciano"))
sexo<-factor(sexo,labels=c("masculino","femenino"))
esco<-factor(esco,labels=c("primaria","secundaria",
             "universitaria","otra"))
econ<-factor(econ,labels=c("bajo","medio","alto"))
vvda<-factor(vvda,labels=c("propietario",
                           "no propietario"))
# marco de datos
datos<-data.frame(edad,sexo,esco,vvda,econ)
```

Análisis de correspondencias múltiples usando la librería `ca`. La opción `nd=NA` indica que se incluyan en la salida el máximo número de dimensiones, si se fija `nd=2`, solo se incluyen dos dimensiones. La opción `supcol=5` indica que la variable suplementaria está en la columna cinco del *data frame*; `lambda='indicator'` es necesario para que el análisis de correspondencias se realice a partir de la tabla disyuntiva.

```
require(ca)
acm<-mjca(datos,lambda="indicator",supcol=5,nd=NA)
summary(acm)
# valores propios
acm$sv
# tabla 3.7
sacm<-summary(acm)
sacm$columns[,c(5,8,6,9)]/1000
# coordenadas de las modalidades
coor<-sacm$columns[,c(5,8)]/1000
# tabla de Burt
acm$Burt
# inercia de las columnas
acm$colinertia
# coordenadas de las columnas
acm$colcoord
# El primer plano factorial
plot(acm)
```

3.7 Análisis de correspondencias múltiples mediante SAS

El procedimiento PROC CORRESP es una rutina computacional del paquete SAS para desarrollar análisis de correspondencias simples o múltiples. El análisis puede hacerse desde una tabla de contingencia, una tabla de Burt o desde los datos categóricos originales.

```
DATA EJEMPLO; /*Archivo de datos Ejemplo*/
INPUT NOMBRE$ EDAD$ GENERO$ ESCOL$ SOCIEC$ VVDA$ @@;
CARDS;
2 1 2 2 1  3 2 1 2 1  3 1 4 2 1  3 2 2 2 1  2 1 1 2 1
1 1 1 1 2  2 1 2 2 2  2 2 2 2 1  3 1 3 3 1  2 2 4 2 1
1 2 3 3 1  1 1 2 3 2  2 2 2 2 1  3 2 2 2 1  3 1 4 3 1
1 2 2 2 1  3 2 3 2 1  3 1 1 1 2  2 2 2 1 2  1 1 3 2 1
;
DATA EJE_MODI;
SET EJEM11_2;/*Nombre de categoría por variable*/
IF EDAD='1' THEN EDAD='JOVEN';
IF EDAD='2' THEN EDAD='ADULTO';
IF EDAD='3' THEN EDAD='VIEJO';
IF GENERO='1' THEN GENERO='HOMBRE';
ELSE GENERO='MUJER';
IF ESCOL='1' THEN ESCOL='PRIMA';
IF ESCOL='2' THEN ESCOL='SECUN';
IF ESCOL='3' THEN ESCOL='UNIVER';
IF  ESCOL='4' THEN ESCOL='OTRO';
IF SOCIEC='1' THEN SOCIEC='BAJO';
IF SOCIEC='2' THEN SOCIEC='MEDIO';
IF SOCIEC='3' THEN SOCIEC='ALTO';
IF VVDA='1' THEN VVDA='PROPIE';  ELSE VVDA='NOPRO';
PROC CORRESP DATA=ACM1 OUTC=EJES OBSERVED MCA;
/*Procedimiento para el ACM,*/
/*EJES contiene las coordenadas de las modalidades
  de variables las activas y suplementarias*/
/*OBSERVED imprime tabla de contingencia*/
/* MCA indica análisis de correspondencias múltiples*/
TABLES EDAD GENERO ESCOL SOCIEC VVDA;
/*TABLES crea una tabla de contingencia o de Burt
```

```
  desde la variables   dadas en el INPUT*/
SUPPLEMENTARY SOCIEC;
/*indica la(s) variables suplementaria(s)*/
DATA EJES1;
SET EJES;
Y=DIM2; X=DIM1; XSYS ='2'; YSYS = '2';
TEXT =_NAME_; SIZE =2; LABEL Y='FACTOR 2'
X='FACTOR 2'; KEEP X Y TEXT XSYS YSYS
SIZE;
PROC GPLOT DATA=EJES1;
SYMBOL V=NONE;
AXIS1 LENGTH=8 IN ORDER=-2 TO 2 BY 0.5;
PLOT Y*X=1/ANNOTATE=EJES1 FRAME HAXIS=AXIS1
VAXIS=AXIS1 HREF=0 VREF=0;
/*Rutina para ubicar las modalidades en
  el primer plano factorial*/
RUN;
```

3.8 Ejercicios

1. El marco de datos `suicide` de la librería `faraway` de R[2] contiene los datos de un año de suicidios en el Reino Unido clasificados por sexo, edad y método.

   a) Colapse el sexo y la edad de los sujetos en un factor simple de seis niveles que contiene todas las combinaciones de sexo y edad. Conduzca un análisis de correspondencia y dé una interpretación del gráfico.

   b) Repita el análisis de correspondencia separadamente para hombres y mujeres. ¿Revela este análisis algo nuevo comparado con el análisis combinado del punto anterior?

2. La tabla 3.8 muestra los datos de 538 pacientes que fueron clasificados en función de 4 tipologías de la enfermedad de Hodgkin (LP, NS, MC, LD) y su respuesta a un tratamiento (Positivo, Parcial, Nulo) al cabo de tres meses. Conduzca un análisis de

[2]Para acceder a los datos debe tener instalada la librería y ejecutar `library(faraway); data(suicide)`.

Tabla 3.8: Respuesta de la enfermedad de Hodgkin a un tratamiento según la tipología.

	Respuesta		
Tipología	Positiva	Parcial	Nula
LP	74	18	12
NS	68	16	12
MC	154	54	58
LD	18	10	44

correspondencias y discuta si el tratamiento actúa igual en todas las tipologías.

Capítulo 4

Modelos log–lineales

4.1 Introducción

En los capítulos anteriores se indagó acerca de la independencia y de algunas medidas de asociación entre las variables que conforman una tabla de contingencia. La mayoría de las investigaciones en las ciencias humanas involucran grandes tablas multidimensionales. En estas tablas se mide un número tal de variables que la interrelación entre ellas resulta compleja. Frecuentemente, el investigador se enfrenta a tablas de dimensión alta. El tratamiento ordinario que hace a los datos contenidos en estas es seleccionar de a dos variables, colapsando las demás. Aunque estos procedimientos pueden ser útiles, tienen los siguiente inconvenientes:

- Se pierde habilidad para estudiar las interacciones entre más de dos variables.
- Las asociaciones marginales entre dos variables es muy distinta de la asociación que podría presentarse con la presencia de otras variables.
- No se pueden estudiar simultáneamente todos los pares de variables.

Cuando se estudian varias mediciones simultáneamente, resulta útil conseguir una descripción parsimoniosa de los datos a través de un modelo

matemático que explique, de alguna forma, las observaciones; a esto se le denomina *modelamiento*. El modelamiento consiste en la aplicación de una serie de procesos (estimación, contrastes de simplificación, diagnosis y replanteamiento) con el objeto de conseguir una explicación apropiada del comportamiento de una *variable respuesta* (*datos*) a partir de una función ponderada de una o más *variables explicativas* (*modelo*). La explicación en general no suele ser perfecta; por tanto, es pertinente observar la discrepancia entre los datos y el modelo. A esta diferencia se le denomina *error* o *residual*. La siguiente igualdad ilustra el concepto de modelo.

$$DATOS \quad = \quad MODELO \quad + \quad ERROR \tag{4.1}$$

La igualdad (4.1) señala que los datos observados, objeto del modelado, son función de un componente *estructural* (sistemático), representado por algún modelo teórico apropiado y un componente *aleatorio* que representa la discrepancia o error entre los datos empíricos y el modelo teórico propuesto. Uno de los ejemplos clásicos de modelos estadísticos es la regresión lineal múltiple, en la cual una variable respuesta (y) es "explicada" a través de unas variables $x_1,\ x_2, \cdots, x_p$ mediante el modelo estadístico

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon = x_i'\beta \tag{4.1a}$$

donde, $x_i' = (1, x_1, \ldots, x_p)$, es el vector de variables explicativas y $\beta' = (\beta_0, \ldots, \beta_p)$, es el vector de parámetros (efectos de las covariables sobre la variable respuesta).

El propósito del modelamiento estadístico es la *búsqueda del modelo más simple que sea capaz de explicar los datos con el mínimo error posible*. Esto equivale a buscar un modelo parsimonioso que se ajuste adecuadamente a los datos. Para el investigador, modelar un conjunto de datos le implica:

1. La *selección* de un modelo teórico simple que, sobre su propia intuición o en el conocimiento sustantivo del objeto de su investigación, contiene una función ponderada de las variables explicativas; esta función puede ser lineal o no lineal en estas variables.

2. La *estimación* de los parámetros a través de la información disponible.

3. El cálculo de una medida de discrepancia para evaluar el ajuste entre los valores observados y los valores pronosticados por el modelo estimado.

4. De acuerdo con la evaluación observada en el numeral anterior, se opta por alguna de las siguientes alternativas:

   - Si el ajuste es apropiado, pero el modelo es complejo, acepta el modelo temporalmente y retorna al comienzo del proceso (1.) en busca de un modelo más simple (parsimonioso).
   - Si el ajuste es adecuado, acepta el modelo de manera tentativa, lo interpreta e integra al marco conceptual donde están inscritos los datos.
   - Si el ajuste no es adecuado, lo rechaza y retorna al comienzo del proceso en busca de un modelo más apropiado (partiendo de la veracidad del marco conceptual asumido por el investigador).

Los modelos para los datos contenidos en una tabla de contingencia son similares a los usados para datos cuantitativos; por ejemplo, los modelos empleados en el *análisis de varianza*. Una consecuencia de esta similitud es que el término *interacción* se usa como una alternativa para describir y cuantificar la asociación entre las variables cualitativas que forman una tabla de contingencia. Cuando se quiera hacer referencia a la interacción entre un par de variables se hablará de interacción de *primer orden*; para la interacción entre tres variables se le dirá interacción de segundo orden, etc.

Uno de los principales atractivos del modelamiento de tablas de contingencia es que los modelos suministran una aproximación sistemática al análisis de tablas complejas y que proveen una estimación de la magnitud de los efectos de interés, en consecuencia, suministran la importancia de diferentes efectos a juzgar (Everitt 1994, 74).

4.1.1 El modelo lineal generalizado

Los modelos lineales generalizados incluyen modelos como los que se expresan en la igualdad (4.1a) (modelo lineal general), modelos logísticos (capítulo 5), modelos de regresión Poisson (capítulo 8), modelos log–lineales para datos multinomiales, entre otros.

Un modelo lineal generalizado se construye con los siguientes tres componentes:

- Una variable respuesta $\{y_i\}$ con alguna distribución de probabilidad (componente aleatoria), $i = 1, 2, \ldots, n$.
- Un conjunto de variables explicativas x_i y un vector de parámetros $\boldsymbol{\beta}$.
- Una función de *enlace*, $g(\dot{})$, entre la componente aleatoria y el componente sistemático μ_i, la cual describe cómo se relaciona con $x_i'\beta$ el valor esperado[1] de Y_i:

$$g(\mu_i) = x_i'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{4.1b}$$

El modelo lineal generalizado se puede construir escogiendo adecuadamente la función de enlace g y la distribución de probabilidad. En el modelo lineal clásico, la distribución de probabilidad es la normal y la función de enlace es la identidad $g(\mu) = \mu$. Para el modelo de regresión logística, la distribución es la binomial y la función de enlace es la logit():

$$g(\mu) = \ln\Big(\frac{\mu}{1-\mu}\Big)$$

Para el modelo de regresión de Poisson, la distribución es la de Poisson y la función de enlace es $g(\mu) = \ln(\mu)$; estas características también son compartidas por los modelos log–lineales. Para una discusión más amplia sobre este tema, se puede consultar Agresti (1996), Agresti (2000) o Dobson (2002).

En el paquete SAS se cuenta con el procedimiento GENMOD para hacer ajuste y diagnóstico sobre este tipo de modelos.

4.2 Modelos log–lineales para tablas de contingencia

4.2.1 El modelo log–lineal

Se inicia la presentación de estos modelos para tablas bidimensionales $f \times c$. La hipótesis de independencia, es decir, de no interacción de primer

[1] $\mathrm{E}(Y_i) = \mu_i$

orden entre las dos variables que conforman la tabla, como se muestra en la sección (2.4) (ecuación 2.11), señala que las probabilidades de cada celda π_{ij} se determinan por las probabilidades marginales fila y columna, es decir,

$$\pi_{ij} = \pi_{i.}\pi_{.j};\ i = 1, 2, \cdots, f, \quad j = 1, 2, \cdots, c. \tag{4.2}$$

Esta igualdad especifica una estructura particular de los datos. Similar a la ecuación (2.12), el valor esperado de las frecuencias es

$$F_{ij} = N\pi_{ij} = N\pi_{i.}\pi_{.j}, \tag{4.3}$$

por tratarse de una distribución multinomial. El modelo log–lineal usa F_{ij} en lugar de π_{ij}; también puede aplicarse en un muestreo tipo Poisson para conteos de celdas. Las frecuencias marginales esperadas son $F_{i.} = N\pi_{i.}$ y $F_{.j} = N\pi_{.j}$. Expresando $\pi_{i.}$ y $\pi_{.j}$ en términos de las frecuencias marginales esperadas y aplicando logaritmo natural en los dos miembros de la ecuación (4.3), se obtiene:

$$\ln F_{ij} = \ln F_{i.} + \ln F_{.j} - \ln N \tag{4.4}$$

Después de sumar sobre i, j e $i-j$, y de algunas simplificaciones, resulta que el modelo expresado en la ecuación (4.3) es equivalente a

$$\ln F_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}, \tag{4.5}$$

donde

$$\mu = \frac{\sum_{i=1}^{f}\sum_{j=1}^{c}\ln F_{ij}}{fc} \tag{4.6a}$$

$$\mu_{1(i)} = \frac{\sum_{j=1}^{c}\ln F_{ij}}{c} - \mu \tag{4.6b}$$

$$\mu_{2(j)} = \frac{\sum_{i=1}^{f}\ln F_{ij}}{f} - \mu \tag{4.6c}$$

El modelo contenido en la igualdad (4.5) se conoce como modelo *log–lineal para independencia*, pues se construye sobre el supuesto de independencia entre las variables fila y columna de una tabla de contingencia. Este modelo es equivalente al modelo multiplicativo $F_{ij} = e^{\mu}e^{\mu_{1(i)}}e^{\mu_{2(j)}}$.

Note la similitud del modelo log–lineal con el modelo para análisis de varianza; de igual manera, el parámetro μ representa el *efecto de la media global*, $\mu_{1(i)}$ representa el *efecto principal* de la i-ésima categoría

de la variable 1 (fila) y $\mu_{2(j)}$ representa el *efecto principal* de la j-ésima categoría de la variable 2 (columna). El modelo dado en la ecuación (4.5) no establece diferencia entre la variable respuesta y las variables explicativas de clasificación.

Una inspección a las ecuaciones (4.6b) y (4.6c) permite apreciar que, en semejanza con las sumas de cuadrados en el análisis de varianza, estas representan los desvíos de la media en fila o en columna del logaritmo de las frecuencias esperadas respecto a la media global μ; en consecuencia:

$$\sum_{i=1}^{f} \mu_{1(i)} = 0 \quad \text{y} \quad \sum_{j=1}^{c} \mu_{2(j)} = 0. \tag{4.7}$$

El modelo anterior está propuesto en términos de las frecuencias teóricas esperadas F_{ij}, que a su vez dependen de unas probabilidades $\pi_{i.}$ y $\pi_{.j}$, también valores desconocidos en la población objeto de estudio. En consecuencia, el conocimiento de estos valores, y de ahí el conocimiento del modelo, se obtienen mediante la estimación de tales parámetros a partir de los datos observados para una tabla particular. Este tema se trata en la siguiente sección.

Un modelo, que extiende el anterior, es el que no parte del supuesto de independencia entre las variables de la tabla de contingencia. De esta manera, se introduce un término que representa la *interacción* entre las variables; el modelo que resulta es el siguiente:

$$\ln F_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)} \tag{4.8}$$

El término $\mu_{12(ij)}$ representa la asociación o interacción entre los niveles i y j de las variables 1 y 2, respectivamente. Si $\mu_{12(ij)} = 0$, para todo i y j, el modelo log–lineal representa el caso de independencia entre las variables 1 y 2. Este modelo tiene el máximo número de parámetros, es decir, tantos parámetros como celdas tenga la tabla; a este tipo de modelos se les denomina *saturados*.

Para tablas de contingencia 2×2, existe una relación directa entre la razón de *odds* (ecuación (2.46c)) y los parámetros de asociación $\mu_{12(ij)}$.

La relación es la siguiente:

$$\begin{aligned}\ln(RO) &= \ln\Big(\frac{F_{11}F_{22}}{F_{12}F_{21}}\Big) = \ln F_{11} + \ln F_{22} - \ln F_{12} - \ln F_{21} \\ &= (\mu + \mu_{1(1)} + \mu_{2(1)} + \mu_{12(11)}) + (\mu + \mu_{1(2)} + \mu_{2(2)} + \mu_{12(22)}) \\ &- (\mu + \mu_{1(1)} + \mu_{2(2)} + \mu_{12(12)}) - (\mu + \mu_{1(2)} + \mu_{2(1)} + \mu_{12(21)}) \\ &= (\mu + \mu_{12(11)} + \mu_{12(22)} - \mu_{12(12)} - \mu_{12(21)})\end{aligned}$$

Así, los $\mu_{12(ij)}$ determinan el logaritmo de la razón de *odds*. Cuando estos parámetros son nulos, el logaritmo de la razón de *odds* es igual a cero, en consecuencia (por el antilogaritmo), la razón de *odds* es igual a 1.0. Por tanto, las variables categóricas (1 y 2) son independientes.

A manera de convenio en la nominación, se hablará de parámetros de interacción de *primer orden* cuando se refieren a la interacción de dos variables, para el caso de tres variables, se nombrará como de *segundo orden*, y así sucesivamente.

El modelo expresado por (4.8) establece que las frecuencias esperadas representan una mezcla de efectos marginales de las variables fila y columna, así como la asociación mutua. La representación de este modelo es equivalente a $F_{ij} = e^{\mu} e^{\mu_{1(i)}} e^{\mu_{2(i)}} e^{\mu_{12(ij)}}$.

El valor de $\mu_{12(ij)}$ se calcula mediante

$$\mu_{12(ij)} = \ln F_{ij} - \mu - \mu_{1(i)} - \mu_{2(j)} \tag{4.9}$$

De acuerdo con la expresión anterior, se tiene la siguiente restricción para los parámetros $\mu_{12(ij)}$:

$$\sum_{j=1}^{c} \mu_{12(ij)} = 0 \quad \text{y} \quad \sum_{i=1}^{f} \mu_{12(ij)} = 0. \tag{4.10}$$

Esta restricción muestra que solo hay $(f-1)(c-1)$ términos linealmente independientes, lo cual es consistente con los grados de libertad asociados a una tabla de contingencia $f \times c$.

La estimación de los parámetros de interacción es útil para identificar las categorías "responsables" de la no independencia entre las variables.

Para el caso de tablas de contingencia en tres variables (tablas $f \times c \times d$), el modelo log–lineal saturado puede escribirse en la forma

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(i)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)} \tag{4.11}$$

El modelo contiene los parámetros $\mu_{1(i)}$, $\mu_{2(i)}$ y $\mu_{3(k)}$, asociados a los efectos principales para cada una de las variables; los parámetros de efectos, de primero ($\mu_{12(ij)}$, $\mu_{13(ik)}$ y $\mu_{23(jk)}$) y segundo orden ($\mu_{123(ijk)}$), de la interacción. Los parámetros de interacción de primer orden representan la interacción o asociación parcial.

4.2.2 Modelos jerárquicos

En este texto se hace la restricción a los llamados *modelos jerárquicos*. En estos modelos la presencia de un parámetro de interacción de orden superior implica la presencia de los parámetros de bajo orden contenidos en tal interacción. Así, por ejemplo, si en el modelo aparece el parámetro μ_{123}, deben también aparecer los parámetros μ_{12}, μ_{13}, μ_{23}, μ_1, μ_2 μ_3. Un modelo como el siguiente

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{3(k)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)}. \qquad (4.12)$$

no se considera jerárquico, pues no aparecen los parámetros μ_2 y μ_{12}, subordinados al parámetro μ_{123}.

La restricción a modelos jerárquicos se debe a problemas de cálculo en la estimación vía máxima verosimilitud (sección 1.5.1). Una explicación más amplia de aspectos, que se sale de los propósitos del curso, se puede consultar en Christensen (1990).

Este tipo de modelos se corresponden con varias hipótesis asociadas a las tablas multidimensionales discutidas en el capítulo 2. Así, la hipótesis de mutua independencia entre las tres variables especifica que no hay interacciones de los dos primeros órdenes[2], es decir,

$$H_0 : \ \mu_{12} = \mu_{13} = \mu_{23} = \mu_{123} = 0. \qquad (4.13)$$

Bajo la hipótesis anterior, el modelo es $\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(i)} + \mu_{3(k)}$, es decir, un modelo de efectos principales. Si un modelo como el último se ajusta adecuadamente a los datos, esto implica que las diferencias entre las frecuencias de las celdas se reflejan por las diferencias entre los marginales simples de cada variable.

En el caso de alguna de las independencias marginales, por ejemplo $\mu_{23} = 0$, para que el modelo sea jerárquico se debe tener que $\mu_{123} = 0$.

[2]Semejante al concepto de independencia mutua entre k-eventos.

El modelo log–lineal resultante es

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{(13ik)} \tag{4.14}$$

El modelo (4.14) especifica que no hay interacción entre las variables 2 y 3 en cada nivel de la variable 1. De otra manera, las variables 2 y 3 son condicionalmente independientes dada la variable 1. En tales modelos, se asume que cada una de las variables 2 y 3 está asociada a la variable 1, puesto que ninguno, μ_{12} y μ_{13}, se han asumido nulos.

Finalmente, la independencia entre la variable 1 respecto a las variables 2 y 3 conjuntas, equivale a que $\mu_{12} = \mu_{13} = \mu_{123} = 0$, de donde se obtiene el modelo log–lineal

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{23(jk)}. \tag{4.15}$$

Observaciones

- El modelo log–lineal puede considerarse un modelo lineal generalizado en el cual el componente sistemático solo admite variables categóricas (nominales u ordinales), el componente de error se distribuye según Poisson y la función de enlace es la transformación logarítmica.
- En el modelo log–lineal no se distingue(n) la(s) variable(s) explicativa(s) de la variable respuesta.
- A diferencia del modelo lineal clásico, en el modelo log–lineal el vector de valores ajustado y el vector de predictores lineales se miden en escalas diferentes.

4.2.3 Estimación de modelos log–lineales

Una de las ventajas de los modelos log–lineales es la estimabilidad de los parámetros, con los cuales se pueden cuantificar los efectos de varias variables y de las interacciones entre estas. La estimación de los parámetros se hace con los logaritmos de los E_{ij} (ecuación 2.14 de la sección 2.4). La forma de estos estimadores es semejante a la que se encuentra para los parámetros en el análisis de varianza. A continuación se ilustra el proceso de estimación mediante los datos contenidos en la tabla 4.1. Los datos corresponden a un estudio realizado sobre 400 pacientes con

Tabla 4.1: Datos de melanoma maligno.

	Sitio del tumor			Total
Tipo de tumor	Cab. & cue.	Tronco	Extrem.	$n_{i.}$
"Pecas" de Hutch	22	2	10	34
Melanoma superf.	16	54	115	185
Nodular	19	33	73	125
Indeterminado	11	17	28	56
Total $n_{.j}$	68	106	226	N=400

Fuente: Stokes et al. (1997, 435).

melanoma maligno. Para cada paciente se registró el sitio del tumor y su característica histológica.

Asumiendo independencia entre las variables tipo de melanoma y el sitio de este, el modelo es:

$$\ln F_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} \text{ con } i = 1, 2, 3, 4;\ j = 1, 2, 3 \qquad (4.16)$$

Se estiman los parámetros del modelo (4.16) mediante el procedimiento de *máxima verosimilitud*, como se expresa en las ecuaciones (4.6a), (4.6b) y (4.6c). Los valores esperados E_{ij} se muestran en la tabla 4.2. La es-

Tabla 4.2: Valores esperados (E^*_{ij}) de los datos de la tabla 4.1.

	Sitio del tumor			Total
Tipo de tumor	Cab. & cue.	Tronco	Extrem.	$n_{i.}$
"Pecas" de Hutch	5.78	9.01	19.21	34
Melanoma superf.	31.45	49.025	104.525	185
Nodular	21.25	33.125	70.625	125
Indeterminado	9.52	14.84	31.64	56
Total $n_{.j}$	68	106	226	N=400

Fuente: Stokes et al. (1997, 435).
**: $E_{ij} = \frac{n_{i.} n_{.j}}{N}$ (Ecuación 2.14).*

timación de la media μ se encuentra mediante la ecuación (4.6a), reem-

plazando F_{ij} por E_{ij}. Así,

$$\begin{aligned}\widehat{\mu} &= \frac{\sum_{i=1}^{f}\sum_{j=1}^{c}\ln E_{ij}}{fc} \\ &= \frac{\ln(5.78)+\ln(9.01)+\cdots+\ln(31.64)}{4\times 3} \\ &= \frac{38.11749796}{12} \\ &= 3.1765\end{aligned}$$

Para los parámetros de efectos principales, los estimadores se consiguen de acuerdo con (4.6b) o (4.6c), por ejemplo, para $\widehat{\mu}_{1(1)}$ es

$$\begin{aligned}\widehat{\mu}_{1(1)} &= \frac{\sum_{j=1}^{c}\ln F_{ij}}{c} - \widehat{\mu} \\ &= \frac{\ln(5.78)+\ln(9.01)+\ln(19.21)}{3} - 3.1764 \\ &= \frac{1.7544+2.1983+2.9554}{3} - 3.1764 \\ &= 2.3027 - 3.1764 \\ &= -0.8737\end{aligned}$$

Mediante el procedimiento CATMOD del paquete SAS se obtienen las estimaciones para los modelos log–lineales. La tabla 4.3 contiene los resultados de la estimación realizada con los datos de la tabla 4.1 para el modelo contenido en la igualdad (4.16). Note que no aparecen las

Tabla 4.3: Parámetros estimados (PROC CATMOD).

Effect	Parameter	Estimate	Standard Error	Chi–Square	Prob
TUMOR	$\widehat{\mu}_{1(1)}$	-0.8737	0.1360	41.28	0.0000
	$\widehat{\mu}_{1(2)}$	0.8203	0.0806	103.60	0.0000
	$\widehat{\mu}_{1(3)}$	0.4282	0.0883	23.53	0.0000
SITIO	$\widehat{\mu}_{2(1)}$	-0.5483	0.0899	37.23	0.0000
	$\widehat{\mu}_{2(2)}$	-0.1044	0.0795	1.72	0.1891

estimaciones para los parámetros $\widehat{\mu}_{1(4)}$ y $\widehat{\mu}_{2(3)}$, pues estos suman cero con los estimadores de las otras modalidades en la misma variable. Así, $\widehat{\mu}_{1(1)}+\widehat{\mu}_{1(2)}+\widehat{\mu}_{1(3)}+\widehat{\mu}_{1(4)}=0$, de donde $\widehat{\mu}_{1(4)}=0-(-0.8737+0.8203+$

$0.4282) = -0.3748$. Similarmente, $\widehat{\mu}_{2(3)} = 0.6527$. La magnitud de estos estimadores es acorde con el tamaño de los totales marginales. Así, en la primera variable (tipo de tumor), el estimador, $\widehat{\mu}_{1(2)}$, de la segunda categoría (tumor tipo superficial) se corresponde con el total marginal más grande. Para la segunda variable, el estimador $\widehat{\mu}_{2(3)}$ de la tercera categoría refleja que esta tiene el total marginal más grande.

Observación
El número de *grados de libertad* es igual al número de celdas en la tabla de contingencia menos el número de parámetros que se deben estimar en el modelo. Para el caso de un modelo saturado de una tabla bidimensional, el número de grados de libertad decrece en la medida que el modelo se hace más complejo. Un modelo saturado tiene 0 grados de libertad; por ejemplo, en el caso bidimensional, $gl = f \times c - [1 + (f-1) + (c-1) + (f-1)(c-1)] = 0$.

4.2.4 Ajuste de los modelos log–lineales

Una vez propuesto un modelo log–lineal y estimadas las frecuencias teóricas, vía máxima verosimilitud, el siguiente paso lógico es medir qué tan bien se ajustan las frecuencias esperadas y las observadas. El propósito de verificar el ajuste de un modelo log–lineal es proveer un soporte para describir y hacer inferencias acerca de la verdadera estructura de asociación entre un conjunto de variables categóricas. Varios modelos log–lineales tienen fórmulas explícitas para estimar las frecuencias esperadas F_{ijk} en términos de los conteos por celda; estos estimadores se llaman *directos* (es el caso de las primeras estimaciones presentadas en la sección 4.2.1). No obstante, muchos modelos log–lineales no tienen estimadores directos. La estimación vía máxima verosimilitud requiere métodos numéricos iterativos para resolver las ecuaciones que de ellos se derivan; uno de estos procedimientos es el algoritmo de *Newton-Raphson*[3]. En la práctica no es relevante conocer cuáles son los modelos que tienen estimación directa, pues la mayor parte de las herramientas computacionales disponibles emplean procedimientos de cálculo que se desarrollan en ambos tipos de modelos.

[3] Agresti (2000, 187–188).

4.2.5 Estadística ji–cuadrado de bondad de ajuste

Sin pérdida de generalidad se explican estas estadísticas para el caso de una tabla de triple entrada. Asuma la hipótesis que las frecuencias esperadas para una tabla tridimensional satisfacen un determinado modelo log–lineal. Como se indica en el capítulo 2, las estadísticas de razón de verosimilitud y ji–cuadrado (para muestras grandes) dan cuenta del *ajuste* mediante la comparación de los valores ajustados y los observados en las celdas de la tabla. De acuerdo con las expresiones (2.15b) y (2.17), estas son, respectivamente,

$$\chi_0^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}, \quad \text{y} \quad G^2 = 2\sum_{i=1}^{f}\sum_{j=1}^{c} n_{ij}\ln\Big(\frac{n_{ij}}{E_{ij}}\Big)$$

Como se anotó, el número de grados de libertad es igual al número de celdas menos el número de parámetros no redundantes –recuerde las restricciones implicadas, por ejemplo en (4.7) y en (4.10)–. Para determinado número de grados de libertad, valores grandes de χ_0^2 o de G^2 tienen p valores pequeños, los cuales son señal de una falta de ajuste del modelo a los datos.

Para contrastar la hipótesis respecto a la significancia de un parámetro, $H_0: \ \mu_{.(\cdot)} = 0$, se emplea la estadística

$$Z = \frac{\widehat{\mu}_{.(\cdot)}}{ASE} \tag{4.17}$$

donde ASE es el error estándar asintótico del parámetro $\widehat{\mu}_{.(\cdot)}$. La estadística Z, bajo H_0, tiene aproximadamente distribución normal estándar; esta se conoce como *estadística de Wald*. La estadística Z^2 tiene distribución ji–cuadrado con un grado de libertad.

Para las estimaciones asociadas a los parámetros del modelo (4.16), las dos últimas columnas de la tabla 4.3, encabezada con "chi-square" y "Prob.", contienen el cuadrado de la estadística de Wald y los p-valores para cada uno de los parámetros estimados, respectivamente. Por ejemplo, en la primera línea de la tabla 4.3 la estimación del parámetro $\mu_{1(1)}$ (tumor tipo Hutchinson) es igual a -0.8737 y el valor del error estándar asintótico (ASE) del parámetro es 0.1360, en consecuencia, el valor de la estadística de Wald es:

$$Z^2 = (-0.8737/0.1360)^2 = 41.28$$

Se advierte la no significancia del parámetro asociado a la categoría "tronco" de la variable sitio ($\mu_{2(2)}$), pues su valor es $p = 0.1891$.

Una alternativa, para observar el ajuste del modelo es el *análisis de varianza* para este tipo de datos, el cual se desarrolla de acuerdo con el modelo asumido (de independencia o no). El procedimiento CATMOD suministra la tabla de análisis de varianza para este modelo; esta se muestra en la tabla 4.4.

Tabla 4.4: Tabla de análisis de varianza (PROC CATMOD).

Source	DF	Chi-Square	Prob
TUMOR	3	121.48	0.0000
SITIO	2	93.30	0.0000
LIKELIHOOD RATIO	6	51.80	0.0000

La tabla de análisis de varianza suministra una fuerte evidencia de que el tipo de tumor y su sitio no son independientes ($G^2 = 51.80$, con 6 g.l. y un $p < 0.0001$). La inquietud ahora se relaciona con el tipo y el sitio del tumor que más altamente asociados se encuentren. Naturalmente, esto implica asumir un modelo que involucre el parámetro de interacción $\mu_{12(ij)}$; este es de la forma

$$\ln F_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}, \quad \text{con } i = 1, 2, 3, 4; \quad j = 1, 2, 3 \tag{4.18}$$

La estimación de los parámetros no es tan directa como las del modelo (4.16), pues las ecuaciones de máxima verosimilitud no tienen una solución explícita como en el caso anterior. Como alternativa se han desarrollado algunos métodos numéricos para encontrar tales estimadores (Agresti 2000, Cap. 6). Dentro del procedimiento CATMOD del paquete SAS se han incorporado estos procedimientos, con los cuales se obtienen estimaciones aproximadas (iterativas) de los parámetros asociados a los modelos log–lineales. A continuación se muestra la sintaxis y la salida del procedimiento aplicado sobre la tabla 4.1; los comandos SAS se escriben en mayúsculas y dentro de los símbolos /∗ y ∗/ se hacen algunos comentarios sobre la sintaxis.

```
DATA  tabla4_1;
INPUT tumor$ sitio $ frec  @@;
CARDS;
Hutchi  CCuello 22 Hutchi  Tronco 2  Hutchi  Extrem 10
Superfi CCuello 16 Superfi Tronco 54 Superfi Extrem 115
Nodular CCuello 19 Nodular Tronco 33 Nodular Extrem 73
Indef   CCuello 11 Indef   Tronco 17 Indef   Extrem 28
;
PROC  CATMOD  ORDER=DATA;
/*invoca el procedimiento CATMOD.*/
/* La instrucción DATA=ORDER, admite las modalidades
   en el orden ingresado*/
WEIGHT  frec;
/*WEIGHT indica que la variable frec corresponde a las
  frecuencias de cada celda*/
MODEL  tumor*sitio= _RESPONSE_;
/*MODEL indica el modelo, en este caso un modelo para
  una tabla bidimensional*/
LOGLIN  tumor  sitio  tumor*sitio;
/*LOGLIN tumor*sitio=_RESPONSE_ indica que se trata de
  un modelo log--lineal*/
/*tumor  sitio  tumor*sitio para señalar que es un
modelo como el que se muestra en la igualdad 4.18*/
RUN;
```

Los resultados del procedimiento anterior se muestran en la tabla 4.5. Se observa que la asociación más alta se presenta entre tumor tipo Hutchinson y la cabeza y el cuello, y que los otros tipos de tumores están más asociados con las extremidades.

4.2.6 Residuales

Tal como se procede en modelos de regresión, los residuales ayudan a mostrar la calidad del ajuste de un modelo log–lineal a los datos de una tabla de contingencia. Los residuales resultan de dividir la diferencia entre los valores observados y ajustados por sus errores estándar. Cuando el modelo se ajusta a los datos, los residuales ajustados tienen aproximadamente una distribución normal estándar. En consecuencia, los residuales cuyo valor absoluto sea superior a 2.0 advierten sobre la

Tabla 4.5: Parámetros estimados para el modelo (4.18).

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
TUMOR	$\widehat{\mu}_{1(1)}$	-1.0451	0.2070	25.49	0.0000
	$\widehat{\mu}_{1(2)}$	0.7615	0.1097	48.16	0.0000
	$\widehat{\mu}_{1(3)}$	0.5031	0.1115	20.37	0.0000
SITIO	$\widehat{\mu}_{2(1)}$	-0.2725	0.1109	6.04	0.0140
	$\widehat{\mu}_{2(2)}$	-0.3210	0.1403	5.23	0.0222
TUM*SIT	$\widehat{\mu}_{12(11)}$	1.3346	0.2359	32.00	0.0000
	$\widehat{\mu}_{12(21)}$	-1.0147	0.3724	7.43	0.0064
	$\widehat{\mu}_{12(31)}$	-0.7904	0.1664	22.56	0.0000
	$\widehat{\mu}_{12(12)}$	0.4745	0.1666	8.11	0.0044
	$\widehat{\mu}_{12(22)}$	-0.3602	0.1626	4.91	0.0267
	$\widehat{\mu}_{12(32)}$	0.2404	0.1735	1.92	0.1659

falta de ajuste del modelo. El cálculo de estos residuales para tablas bidimensionales se hace conforme a la expresión (2.38); para tablas de tamaño superior, Agresti (1990: 224-227) muestra algunas fórmulas para su cómputo. Mediante el procedimiento CATMOD del SAS se obtienen los residuales; con la ayuda de una hoja de cálculo (como Excel) se pueden obtener los residuales ajustados.

Para el caso de una tabla tridimensional, considere los datos de la tabla 4.6 relacionados con enfermedades coronarias. Un grupo de 1.330 pacientes se clasificaron respecto las variables presión sanguínea, colesterol y enfermedad coronaria[4].

Primero se considera el modelo solo con los efectos principales; esto equivale a suponer que las variables son mutuamente independientes. El modelo es

$$\ln F_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} \tag{4.19}$$

El número de grados de libertad de este modelo es

$$\begin{aligned} gl =& f \times c \times d - [1 + (f-1) + (c-1) + (d-1)] \\ =& 4 \times 4 \times 2 - [1 + (4-1) + (4-1) + (2-1)] = 24 \end{aligned}$$

[4]Tomado de Everitt (1992: 87).

Variable	Categorías
(V1). Presión sanguínea	Pre1: menos de 127 mm de Hg
	Pre2: entre 127 y 146
	Pre3: entre 147 y 167
	Pre4: mayor de 147
(V2). Colesterol	Col1: menos de 200 mg/100 cc
	Col2: entre 200 y 219
	Col3: entre 220 y 259
	Col4: mayor de 260
(V3). Enfermedad coronaria	Sí
	No

Tabla 4.6: Datos sobre enfermedades coronarias.

Enf. coronaria	Pre. sanguínea	Colesterol				Total
		Col1	Col2	Col3	Col4	
Sí	Pre1	2	3	3	4	12
	Pre2	3	2	1	3	9
	Pre3	8	11	6	6	31
	Pre4	7	12	11	11	41
Subtotal		20	28	21	24	93
No	Pre1	117	121	47	22	307
	Pre2	85	98	43	20	246
	Pre3	119	209	68	43	439
	Pre4	67	99	46	33	245
Subtotal		388	527	204	118	1237
Total		408	555	225	142	1330

El modelo log–lineal solo con parámetros para efectos principales –modelo (4.19)– no provee un ajuste adecuado a los datos, pues se rechaza la hipótesis de independencia, ya que el valor de la estadística de razón de verosimilitud es $G^2 = 78.96$, con un $p < 0.001$.

El siguiente modelo por considerar contiene parámetros para efectos principales y para interacciones de primer orden entre las tres variables. Este se muestra en la siguiente expresión:

Tabla 4.7: Parámetros estimados para el modelo (4.19).

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
VAR1	$\widehat{\mu}_{1(1)}$	-0.0136	0.0486	0.08	0.7795
	$\widehat{\mu}_{1(2)}$	-0.2375	0.0525	20.49	0.0000
	$\widehat{\mu}_{1(3)}$	0.3739	0.0431	75.32	0.0000
VAR2	$\widehat{\mu}_{2(1)}$	0.3357	0.0470	51.02	0.0000
	$\widehat{\mu}_{2(2)}$	0.6434	0.0434	219.67	0.0000
	$\widehat{\mu}_{2(3)}$	-0.2594	0.0566	21.00	0.0000
VAR3	$\widehat{\mu}_{3(1)}$	1.2939	0.0538	579.36	0.0000

$$\ln F_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} \quad (4.20)$$

con $i = 1, 2, 3, 4$; $j = 1, 2, 3, 4$ y $k = 1, 2$

Los parámetros de este modelo se estiman de manera iterativa mediante el procedimiento CATMOD del paquete SAS. Para estos datos, el modelo (4.20) muestra un ajuste a los datos bastante bueno, pues $G^2 = 4.77$ con un p valor de 0.8535, luego no se requieren parámetros de interacción de segundo orden para ajustar estos datos.

La tabla 4.8 muestra los parámetros estimados para el modelo (4.20) junto con las estadísticas de significancia de los mismos. Note que no es necesario escribir todos los parámetros pues, por la restricción impuesta para la estimación, estos deben sumar cero en cada uno de los subíndices. Por ejemplo, $\mu_{1(1)} + \mu_{1(2)} + \mu_{1(3)} + \mu_{1(4)} = 0$, de donde $\mu_{1(4)} = 0.3116$.

De acuerdo con la significancia de los parámetros mostrada en la tabla 4.8, se sugiere el modelo que incluya solo los parámetros $\mu_{13(\cdot)}$ $\mu_{23(\cdot)}$. El modelo es

$$\ln F_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ik)} + \mu_{23(jk)},$$
$$\text{con } i = 1, 2, 3, 4;\ j = 1, 2, 3, 4 \text{ y } k = 1, 2 \quad (4.21)$$

Este modelo implica asumir que no hay asociación entre la presión

Tabla 4.8: Parámetros estimados para el modelo (4.20).

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
VAR1	$\widehat{\mu}_{1(1)}$	-0.2462	0.1244	3.92	0.0478
	$\widehat{\mu}_{1(2)}$	-0.4793	0.1377	12.11	0.0005*
	$\widehat{\mu}_{1(3)}$	0.4139	0.0936	19.56	0.0000*
VAR2	$\widehat{\mu}_{2(1)}$	0.1675	0.0993	2.85	0.0916
	$\widehat{\mu}_{2(2)}$	0.4312	0.0897	23.10	0.0000*
	$\widehat{\mu}_{2(3)}$	-0.1703	0.1002	2.89	0.0893
VAR3	$\widehat{\mu}_{3(1)}$	-1.3018	0.0652	398.83	0.0000*
VAR1*VAR2	$\widehat{\mu}_{12(11)}$	0.2223	0.0828	7.21	0.0072*
	$\widehat{\mu}_{12(12)}$	-0.0184	0.0804	0.05	0.8194
	$\widehat{\mu}_{12(13)}$	-0.0367	0.1038	0.12	0.7238
	$\widehat{\mu}_{12(21)}$	0.1105	0.0894	1.53	0.2164
	$\widehat{\mu}_{12(22)}$	-0.0436	0.0858	0.26	0.6115
	$\widehat{\mu}_{12(23)}$	0.0274	0.1089	0.06	0.8010
	$\widehat{\mu}_{12(31)}$	-0.1135	0.0761	2.22	0.1361
	$\widehat{\mu}_{12(32)}$	0.1549	0.0684	5.13	0.0235
	$\widehat{\mu}_{12(33)}$	-0.0619	0.0905	0.47	0.4943
VAR1*VAR3	$\widehat{\mu}_{13(11)}$	-0.2267	0.1231	3.39	0.0657
	$\widehat{\mu}_{13(21)}$	-0.2724	0.1369	3.96	0.0466
	$\widehat{\mu}_{13(31)}$	0.0545	0.0928	0.34	0.5573
VAR2*VAR3	$\widehat{\mu}_{23(11)}$	-0.2191	0.0992	4.88	0.0272*
	$\widehat{\mu}_{23(21)}$	-0.2383	0.0889	7.19	0.0073*
	$\widehat{\mu}_{23(31)}$	0.0745	0.0993	0.56	0.4530

**: significativo al* 1%.

sanguínea y el nivel de colesterol tanto para pacientes que padecen enfermedades coronarias como para los que no las padecen. El ajuste del modelo provee un $G^2 = 24.40$ con un p valor de 0.1423, con lo cual parece que el modelo (4.21) se ajusta adecuadamente a los datos. La tabla 4.9 contiene las estimaciones de este modelo de acuerdo con los

datos de la tabla 4.6.

Tabla 4.9: Parámetros estimados para el modelo (4.21).

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
VAR1	$\widehat{\mu}_{1(1)}$	-0.2251	0.1223	3.39	0.0657
	$\widehat{\mu}_{1(2)}$	-0.4797	0.1361	12.42	0.0004*
	$\widehat{\mu}_{1(3)}$	0.4283	0.0919	21.70	0.0000*
VAR2	$\widehat{\mu}_{2(1)}$	0.1199	0.0979	1.50	0.2206
	$\widehat{\mu}_{2(2)}$	0.4413	0.0878	25.24	0.0000*
	$\widehat{\mu}_{2(3)}$	-0.1771	0.0979	3.28	0.0703
VAR3	$\widehat{\mu}_{3(1)}$	-1.3004	0.0653	397.10	0.0000*
VAR1*VAR3	$\widehat{\mu}_{13(11)}$	-0.2474	0.1223	4.09	0.0431
	$\widehat{\mu}_{13(21)}$	-0.2805	0.1361	4.25	0.0393
	$\widehat{\mu}_{13(31)}$	0.0483	0.0919	0.28	0.5995
VAR2*VAR3	$\widehat{\mu}_{23(11)}$	-0.2618	0.0979	7.15	0.0075*
	$\widehat{\mu}_{23(21)}$	-0.2467	0.0878	7.89	0.0050*
	$\widehat{\mu}_{23(31)}$	0.0840	0.0979	0.74	0.3906

*: *significativo al* 1%.

Como conclusión se puede afirmar que existe una asociación positiva entre los niveles altos de presión sanguínea (nivel 4) y la ocurrencia de enfermedades coronarias, pues $\widehat{\mu}_{13(41)} = 0.4796$ con $p < 0.01$. Por un argumento semejante se muestra la asociación positiva entre los niveles de colesterol alto y la presencia de enfermedades coronarias ($\widehat{\mu}_{23(41)} = 0.4245$ con $p < 0.01$).

Además, la no presencia de interacción de segundo orden implica, por una parte, que la asociación entre la presión sanguínea y las enfermedades coronarias es igual en todos los niveles de colesterol y, por otra, que la asociación entre el colesterol y las enfermedades cardiacas es igual para todos los niveles de presión sanguínea.

4.3 Procesamiento de datos con R

A continuación se muestra cómo realizar los cálculos para la estimación e inferencia con `R` mediante los datos contenidos en la tabla 4.1.

```
# introducción de los datos de melanoma maligno.
```

```
 x<-matrix(c(22,16,19,11,
             2,54,33,17,
            10,115,73,28),
            nrow=4)
dimnames(x)<-list(TipoTumor=c("Pecas","M.Superf",
                              "Nodular","Indet"),
           Sitio=c("Cab.Cue","Tronco","Extrem") )
f<-nrow(x)    # número de filas
c<-ncol(x)    # número de columnas
addmargins(x) # la tabla junto con las marginales
```

Para ajustar un modelo log-lineal se deben tener los datos en un marco de datos (*data frame*). A continuación se usa la función `melt` de la librería `reshape` que convierte la tabla `x` en un data frame con dos factores (filas y columnas) y una variable numérica que corresponde a las entradas de la tabla (`value`).

```
require(reshape)
datos<-melt(x) # convierte la tabla x en un data frame
datos
summary(datos)
```

A continuación se ajusta el modelo 4.16, usando la función `glm`.

```
mod1<-glm(value~TipoTumor+Sitio,data=datos,family=
          poisson(link="log"))
summary(mod1)
# valores esperados
esperados<-predict(mod1,type="response")
# tabla 4.2
E<-matrix(esperados,nrow=4)
dimnames(E)<-dimnames(x)
addmargins(E)
```

Para obtener las estimaciones de los parámetros junto con sus errores estándar, el estadístico ji–cuadrado y el p–valor, como se muestra en la tabla 4.3, se usa el siguiente código:

```
estimacion<-predict(mod1,type="terms",se.fit=TRUE)
estimados<-c(unique(estimacion$fit[,1]),
             unique(estimacion$fit[,2] ))
# estimaciones
estimados<-round(estimados,4)
ErrorEs<-c(unique(estimacion$se.fit[,1]),
           unique(estimacion$se.fit[,2]))
# errores estándar
ErrorEs<-round(ErrorEs,4)
# valor de ji--cuadrado (redondeado a dos cifras)
Chi<-round( (estimados/ErrorEs)^2, 2)
# pvalor (redondeado a cuatro cifras)
p.valor<-round( pchisq(Chi,1,lower.tail=FALSE),4)
efecto<-rep(c("Tumor","Sitio"),c(f,c) )
# organizados en una tabla
data.frame(efecto,estimados,ErrorEs,Chi,p.valor)
```

La tabla de análisis de varianza (tabla 4.4) se obtiene mediante el código

```
anova(mod1,test="Chisq")
```

Otra forma de probar la hipótesis de independencia es comparando el modelo sin interacción con el modelo que incluye el efecto de la interacción (modelo 4.18), mediante la función anova, como se muestra a continuación:

```
# modelo con interacción (4.18)
mod2<-glm(value~TipoTumor*Sitio,data=datos,
          family=poisson(link = "log"))
# prueba de la hipótesis de independencia
anova(mod1,mod2,test="Chisq")
```

Hay dos formas más de ajustar modelos log–lineales en `R`: mediante la función `loglin()` y mediante la función `loglm()` de la librería `MASS`. Veamos unos ejemplos:

```
# modelo 4.16 (sin interacción) mediante loglin
mlogl1<-loglin(x,margin=list("TipoTumor","Sitio"),
               fit=T,param = T)
```

```
mlogl1$param
# modelo 4.18 (con interacción) mediante loglin
mlogl2<-loglin(x,margin=list(c("TipoTumor","Sitio") ),
               fit=T,param = T)
mlogl2$param
```

El ajuste de los mismos modelos mediante la función `loglm()` de la librería `MASS` es

```
library(MASS)
ajuste1<-loglm(value~TipoTumor+Sitio,data=datos)
# Estadísticos de la razón de verosimilitud y de pearson
# ¿qué tan bueno es el ajuste del modelo?
anova(ajuste1)
# valores ajustados (tabla 4.2)
fitted(ajuste1)
# primera columna da la tabla 4.3
coef(ajuste1)$TipoTumor
coef(ajuste1)$Sitio
# otra forma de ver la prueba de bondad de ajuste
# del modelo.
summary(ajuste1)$test
```

4.4 Ejercicios

1. En un estudio retrospectivo de caso–control, un grupo de pacientes con úlcera fue comparado a un grupo de pacientes de control que no tenían úlcera pero eran similares a los pacientes con úlcera con respecto a la edad, sexo y estrato socioeconómico. Los pacientes fueron clasificados de acuerdo con el sitio de la úlcera: gástrica o duodenal. A todos los sujetos se les preguntó sobre el uso o no de aspirina. Los resultados se muestran en la tabla 4.11. Ajuste modelos log–lineales apropiados con el fin de dar respuesta a algunas preguntas de interés como:

   a) ¿Está la úlcera gástrica asociada al uso de aspirina?

   b) ¿Está la úlcera duodenal asociada al uso de aspirina?

c) Si existe alguna asociación de la úlcera al uso de aspirina, ¿es igual para ambos sitios?

2. La tabla 4.10 muestra datos sobre la relación entre la raza y la imposición de la pena de muerte. Analice los datos.

Tabla 4.10: Raza y pena de muerte.

Raza del defensor	Raza de la víctima	Pena de muerte	
		Sí	No
Negra	Negra	6	97
	Blanca	11	52
Blanca	Negra	0	9
	Blanca	19	132

Tabla 4.11: Úlceras gástrica y duodenal en relación con el uso de aspirina.

	Uso de aspirina		
	No	Sí	Total
Úlcera gástrica			
Control	62	6	68
Casos	39	25	64
Úlcera duodenal			
Control	53	8	61
Casos	49	8	57

3. Considere los datos del ejercicio 1 del capítulo 2. Lleve a cabo la prueba de independencia mediante la comparación del ajuste del modelo sin interacción con el del modelo que incluye todas las interacciones.

4. Considere los datos del ejercicio 2 del capítulo 2. Lleve a cabo la prueba de independencia mediante la comparación del ajuste del modelo sin interacción con el del modelo que incluye todas las interacciones.

Capítulo 5

Regresión logística

5.1 Introducción

Como se expresa en la ecuación (4.1), un modelo estadístico tiene como finalidad principal explicar el comportamiento (en términos de variabilidad) de las variables que, de acuerdo con el marco conceptual asumido por el investigador, están ligadas a un fenómeno mediante otras variables asociadas al mismo fenómeno. Un modelo está compuesto por la variable a explicar (dependiente o respuesta) y las variables explicativas (independientes o regresoras) con las cuales se pretende dar cuenta del comportamiento de la variable respuesta. El modelo se hace visible a través de una función matemática con la cual se expresan las relaciones entre las variables puestas en juego.

A continuación se listan algunos casos que se pueden abordar con la técnica que se desarrollará de manera general en este capítulo.

- Un sujeto operado se infecta o no durante cierto lapso posoperatorio.
- Un bebé nace con malformación congénita o sin esta.
- Un paciente hospitalizado muere o no antes del alta.
- A los tres meses de vida, un niño ha dejado de lactar o aún se mantiene alimentándose con leche materna.

- Un año después de una intervención quirúrgica, se ha resuelto o no el problema que la originó.
- Después de un tratamiento de quimioterapia en un paciente con cáncer de pulmón se observa alguno de los siguientes resultados sobre la enfermedad: aumento, no cambio, remisión parcial, remisión completa.

En casos como los anteriores, usualmente el interés se presta sobre la evaluación del efecto de uno o más antecedentes relacionables[1] con la ocurrencia del evento.

A diferencia del último caso, los demás eventos muestran solo dos resultados: *ocurrencia* o *no ocurrencia* de un hecho. Note por Y la variable dependiente que indica la ocurrencia o no del suceso, esta es dicotómica. Admita que asume los valores

$$Y = 1, \text{ si el hecho ocurre}$$
$$Y = 0, \text{ si el hecho no ocurre}$$

El caso más sencillo trata de evaluar el efecto de un único factor X sobre una variable Y.

5.2 Modelo de regresión logística

A manera de ejemplo, considere el caso que estudia la infección hospitalaria quirúrgica en pacientes intervenidos de la cadera. De manera que $Y = 1$, cuando el paciente se infecta a lo largo de la primera semana, y $Y = 0$, si no se infecta. Se quiere evaluar un nuevo modelo técnico-organizativo del servicio de enfermería que se dispensa a estos pacientes. Sea X_1 una variable dicotómica que vale 0 si el sujeto estuvo tratado bajo el nuevo modelo y vale 1 en caso de que haya sido tratado por el modelo tradicional. Además, se quiere evaluar si la edad del paciente, X_2, se asocia al desarrollo o no de la infección.

[1] Son "relacionables" hasta tanto no se tenga evidencia estadística y conceptual suficiente.

La tabla 5.1 contiene la información sobre 40 pacientes, divididos en dos grupos de 20, cada uno de los cuales estuvo sometido a uno de los dos tipos de atención. Una primera inquietud es si existe asociación entre

Tabla 5.1: Infección en pacientes hospitalizados.

Atención tradicional		Atención propuesta	
Edad	Infección	Edad	Infección
34	No	45	No
21	No	23	No
54	Sí	44	No
67	No	65	Sí
32	Sí	66	Sí
56	Sí	74	Sí
76	Sí	34	No
44	No	43	No
34	No	47	No
21	No	37	No
48	No	26	No
39	No	54	No
22	No	53	No
45	No	55	Sí
65	Sí	23	No
67	Sí	34	No
22	No	43	No
32	No	45	No
21	Sí	31	No
76	Sí	55	No

Fuente: Silva (1995, 4).

el modelo de atención de enfermería y el desarrollo de una infección. Este problema se resuelve como se muestra en el capítulo 2, mediante el análisis de una tabla de contingencia 2×2, mostrada en la tabla 5.2. La tasa de infección entre los cobijados por el modelo de atención propuesto ($4/20 = 0.2$) es la mitad de los que corresponden al modelo tradicional ($8/20 = 0.4$). No obstante, la estadística ji–cuadrado (ecuación 2.15b) muestra un valor de $\chi_0^2 = 1.90$, el cual es menor que 3.84 (el cuantil de una ji–cuadrado con un grado de libertad y $\alpha = 0.05$); por tanto, se

Tabla 5.2: Pacientes por modelo de atención y condición de infección.

	Condición de infección		
Modelo de atención	Infectados	No infectados	Total
Mod. tradicional	8	12	20
Mod. propuesto	4	16	20
Total	12	28	40

puede afirmar que las tasas de infección no difieren significativamente en los dos tipos de tratamientos.

Con relación a la edad, se podría comparar la media de edad de los que se infectaron con la media de la edad de los que no se infectaron. Las medias, de acuerdo con la tabla 5.1, son 58.9 y 38.1 años, respectivamente. Las estadísticas para verificar la igualdad de medias, paramétrica como la t-Student o no paramétrica como la de Wilcoxon (capítulo 7), coinciden en advertir que existe una diferencia significativa entre la edad promedio de los dos grupos.

Ninguna de las dos soluciones anteriores emplea la regresión. Es más, como la intención es evaluar si la adquisición de la infección o no Y es dependiente de los valores asumidos por la variable independiente tenida en cuenta, se puede considerar la variable Y en función de la variable X_1, de X_2 o ambas.

La técnica estadística que se emplea para estos propósitos es la *regresión* lineal. Los modelos de regresión con los que se podría evaluar la adquisición de la infección son:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \epsilon_1, \\ Y &= \beta_0 + \beta_2 X_2 + \epsilon_2 \text{ o} \\ Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_3 \end{aligned} \tag{5.1}$$

Estos modelos tienen los siguientes inconvenientes (Peña, 1998: 501):

1. El valor esperado, evaluado para valores particulares de las variables independientes (en el i-ésimo individuo), corresponde a *la probabilidad de que la característica estudiada esté presente* ($p_i = P(Y = 1|x_{i1}, x_{i2})$). Este será un número, que salvo algunas excepciones, estará entre 0.0 y 1.0.

2. Conocidos los valores de las X, los únicos valores posibles de Y son 0.0 y 1.0; por tanto la distribución de los ϵ_i es discreta, con valores, por ejemplo, para el primer modelo de los contenidos en (5.1), $(1 - \beta_0 + \beta_1 X_1)$ y $-(\beta_0 + \beta_1 X_1)$, es decir, $(1 - p_i)$ y $(-p_i)$. Se verifica que el valor esperado ϵ es

$$\mathrm{E}(\epsilon) = p_i(1 - p_i) + (1 - p_i)(-p_i) = 0 \tag{5.1}$$

por consiguiente, la variable ϵ_i tiene media cero, pero no sigue una distribución normal. La varianza de ϵ_i es

$$\mathrm{var}(\epsilon_i) = (1 - p_i)^2 p_i + (1 - p_i)p_i^2 = (1 - p_i)p_i \tag{5.2}$$

de manera que la varianza de los ϵ_i no es constante (hay heterocedasticidad).

En consecuencia, la regresión lineal debe ser descartada como alternativa a estas situaciones. La opción es la *regresión logística*.

Con la regresión logística se procura expresar *la probabilidad* de que ocurra el evento de interés como función de algunas variables, que desde la teoría (o la experiencia) se asumen como influyentes.

En su forma más simple el modelo logístico incluye una sola variable explicativa, por ejemplo X_1; este es

$$P(Y = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1)]} \tag{5.3}$$

El caso más general, que involucra p-variables explicativas $X_1, \ldots, X_p$, es el siguiente:

$$P(Y = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)]} \tag{5.4}$$

donde $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ son los parámetros del modelo; $\exp(\cdot)$ se refiere a la función exponencial. La expresión del lado derecho de (5.4) se conoce con el nombre de *función logística multivariada*, mientras que (5.3) corresponde a la función logística univariada. Con relación a las variables explicativas del modelo de regresión logística, estas pueden ser de tipo nominal, ordinal o continuo. Este es uno de los grandes atractivos de la regresión logística. La figura (5.1) muestra la función logística univariada.

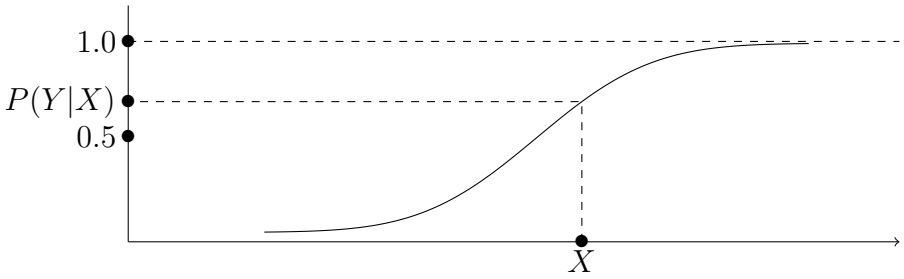

Figura 5.1: Función logística

Para ilustrar los conceptos inherentes con la regresión logística, por ahora, admita que la estimación de los parámetros del modelo

$$P(Y=1) = \frac{1}{1+exp[-(\beta_0+\beta_1 X_1+\beta_2 X_2)]} \tag{5.5}$$

arroja los siguientes resultados[2]:

$$\begin{array}{ll} INTERCEPTO: & \widehat{\beta}_0 = 5.3363 \\ MODELO: & \widehat{\beta}_1 = 1.4357 \\ EDAD: & \widehat{\beta}_2 = -0.1077 \end{array}$$

De manera que para un paciente de 54 años de edad cuya atención de enfermería sea la tradicional (es decir: $X_1 = 0$ y $X_2 = 54$), la probabilidad de contraer infecciones posteriores a la cirugía de cadera se estima remplazando las estimaciones de los β y los valores de X_1 y X_2 en (5.4), de la siguiente manera:

$$\begin{aligned} P(Y=1) &= \frac{1}{1+\exp\left[-(5.3363+1.4357(0)-0.1077(54))\right]} \\ &= \frac{1}{1+\exp\left[-(-0.4795)\right]} \\ &= 0.3824 \end{aligned}$$

Este resultado significa que aproximadamente al 38% de los pacientes con este perfil presentará infecciones en el transcurso de la primera semana poscirugía.

[2]Se obtuvieron mediante el procedimiento LOGISTIC del SAS.

Es importante anotar que la codificación sobre que es $Y = 1$ o que es $Y = 0$ es arbitraria e irrelevante. De forma que si se considera al revés de la optada anteriormente, los parámetros serían los mismos, excepto que el signo sería el opuesto al que tenía antes. Para el ejemplo anterior, si se hace que $Z = 1$ corresponde a no contraer infección y $Z = 0$ a contraer infección, entonces el modelo estimado resulta

$$P(Z = 1) = \frac{1}{1 + \exp\left[-(-5.3363 - 1.4357X_1 + 0.1077X_2)\right]}$$

Para calcular la probabilidad de que un individuo se infecte durante la primera semana después de la cirugía se sigue del hecho que

$$\begin{aligned} P(Z = 1) &= \frac{1}{1 + \exp\left[-(-5.3363 - 1.4357(0) + 0.1077(54))\right]} \\ &= \frac{1}{1 + \exp\left[-(0.4795)\right]} \\ &= 0.6176 \end{aligned}$$

corresponde al complemento de la probabilidad de que $Y = 1$, es decir, que $P(Y = 1) = 1 - P(Z = 1) = P(Z = 0)$.

De este resultado se debe tener presente la manera en que se ha definido la variable de respuesta: pues un coeficiente con el signo positivo indica que $P(Y = 1)$ crece cuando lo hace la variable asociada al respectivo coeficiente, pero el sentido cualitativo de este hecho depende de lo que representen tanto la variable en cuestión como el evento $Y = 1$. En la siguiente sección se ampliará la interpretación de estos coeficientes.

5.3 Interpretación de los coeficientes de regresión

Una interpretación adecuada de los coeficientes β en un modelo de regresión logística va de la mano con los conceptos de: riesgo relativo, *odds* y la de razón *odds*, explicados en las secciones (2.5.5) y (2.5.6). En tal sentido, si se considera que las p-variables conforman un vector X, es decir, que $X = (X_1, X_2, \ldots, X_p)$, se puede probar, por las propiedades de la función exponencial, que los *odds* del evento $Y = 1$ se pueden escribir como

$$O(X) = \frac{P(Y=1)}{P(Y \neq 1)} = \frac{P(Y=1)}{1-P(Y=1)}$$
$$= \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) \tag{5.6}$$

Suponga que se tienen dos perfiles específicos, es decir, dos individuos k y l determinados por los valores que asuman en cada una de las p-variables; estos son:

$$\begin{array}{ll} \text{individuo k} : & X_{k1}, X_{k2}, \ldots, X_{kp} \\ \text{individuo l} : & X_{l1}, X_{l2}, \ldots, X_{lp} \end{array}$$

el valor de los *odds* en cada uno de ellos, de acuerdo con (5.6), son $O(X^k)$ y $O(X^l)$, respectivamente. Así, $O(X^k)$ representa los *odds* correspondientes al primer perfil y $O(X^l)$ los relacionados con el segundo. Mediante manipulación algebraica sencilla se obtiene la siguiente expresión:

$$RR = \frac{O(X^k)}{O(X^l)} = \exp\left[\sum_{i=1}^{p} \beta_i (X_{ki} - X_{li})\right] \tag{5.7}$$

donde X^k y X^l denotan el vector de observaciones para los individuos k y l, respectivamente.

La expresión (5.7) corresponde a una medida relativa del riesgo relacionada con un perfil respecto de otro en términos de los parámetros de la regresión logística.

Para el ejemplo hasta aquí tratado, la fórmula (5.7) permite responder preguntas como: ¿cuánto más riesgo tiene un sujeto de 50 años asistido por la metodología propuesta, que uno de 55 años asistido por la metodología tradicional? En este caso los perfiles son:

$$\begin{array}{ll} \textit{individuo } 1 : & X_{11} = 1 \text{ y } X_{12} = 50 \\ \textit{individuo } 2 : & X_{21} = 0 \text{ y } X_{22} = 55 \end{array}$$

Al remplazar en (5.6) el valor de las variables y las estimaciones de los parámetros, se obtiene:

$$\frac{O(X^1)}{O(X^2)} = \exp[\beta_1(X_{11} - X_{21}) + \beta_2(X_{12} - X_{22})]$$
$$= \exp[\beta_1(1-0) + \beta_2(50-55)]$$
$$= 7.2001$$

Esto quiere decir que la primera situación (descrita por el perfil del individuo 1) es 7 veces más "peligrosa" que la segunda.

Si los perfiles son iguales, salvo en la i-ésima variable, en (5.6) todos los sumandos diferentes al i-ésimo se anulan, de donde resulta:

$$\frac{O(X^k)}{O(X^l)} = \exp[\beta_i(X_{ki} - X_{li})] \tag{5.8}$$

Para el ejemplo, este caso es trivial, puesto que tan solo se tienen dos variables. Suponga que los individuos están expuestos al mismo tratamiento y que las edades son 45 y 60 años.

$$\frac{O(X^k)}{O(X^l)} = \exp[-0.1077(45 - 60)] = 5.0304$$

Esto quiere decir que la primera situación es 5 veces más peligrosa (en términos de contraer infecciones) que la segunda. Aunque parezca una situación extraña, puesto que se esperaría que la metodología de tratamiento propuesta produjera mejores resultados (menos riesgo) que la tradicional, note que tal metodología parece "favorecer" a los pacientes con edad avanzada.

Finalmente, si $X_{ki} = X_{li} + 1$ y todos los demás valores de las otras variables iguales, (5.7) se reduce a:

$$\frac{O(X^k)}{O(X^l)} = \exp(\beta_i) \tag{5.9}$$

Por ejemplo, si los sujetos difieren tan solo en que uno se expuso al tratamiento propuesto y el otro no, entonces:

$$\begin{aligned}\frac{O(X^1)}{O(X^2)} &= \exp(1.4357) \\ &= 4.2026\end{aligned}$$

de donde se concluye que la metodología de asistencia propuesta aumenta el riesgo de infecciones en 4 veces, suponiendo que se han controlado las demás variables.

5.4 Construcción e interpretación de la función logística

Igual que en regresión múltiple, se deben *estimar* los parámetros β a partir de la información registrada sobre n individuos (u objetos). Tal información se dispone en notación matricial así:

$$\begin{pmatrix} Y_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Y_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

Cada fila representa el resultado de medir las variables Y, $X_1, \ldots, X_p$ en un individuo; la primera columna está compuesta por unos y ceros. La matriz tiene tantas filas como sujetos haya en la muestra. Por ejemplo, X_{35} representa la medición de la quinta variable explicativa sobre el tercer individuo de la muestra.

En regresión lineal el método usado con más frecuencia es el de *mínimos cuadrados*. Con el cual se buscan los valores de $\beta_0, \beta_1, \ldots, \beta_p$ que minimicen la suma de cuadrados de las desviaciones entre los valores observados de Y y los valores pronosticados por el modelo, es decir, se intenta encontrar los valores de los parámetros que minimicen el *error de predicción*. Bajo algunos supuestos, como varianza constante, normalidad de los errores, independencia entre las variables explicativas, el método de los mínimos cuadrados produce estimadores con propiedades estadísticas deseables. Pero como se anota en las igualdades (5.1) y (5.2), las variables dicotómicas no reúnen estas propiedades; en consecuencia, se debe optar por otro procedimiento de estimación.

Dado que la variable Y sigue una distribución tipo Bernoulli (sección (1.4.2)), el procedimiento adecuado de estimación es de *máxima verosimilitud* (sección (1.5.1)). La solución a las ecuaciones implicadas en este proceso de optimización suministran los estimadores de los parámetros del modelo, tal explicación se sale de los propósitos de este texto; los interesados pueden consultar a Hosmer & Lemeshov (1989), Agresti (2000) Paulino (2006). No obstante, los paquetes estadísticos como SAS y R tienen programados los algoritmos con los cuales se obtienen las estimaciones de los parámetros.

Considere el caso de personas mayores de 40 años que no han padecido

enfermedad coronaria alguna. Se define la variable Y, la cual registra el desarrollo de alguna enfermedad coronaria durante un periodo de observación de 10 años, así:

$$Y = \begin{cases} 1, & \text{si desarrolló la enfermedad coronaria durante el periodo} \\ 0, & \text{si no apareció la enfermedad coronaria durante el mismo periodo} \end{cases}$$

Además admítase que se consideran las siguientes variables como explicativas del resultado de la variable Y, las cuales se miden al comienzo del período:

X_1 : Edad del individuo (EDAD)
X_2 : tabaquismo: 1 si fumaba, 0 en caso contrario (fuma)
X_3 : tensión arterial sistólica (TAS)

La matriz de datos para este caso está contenida en la tabla 5.3[3]: Por ejemplo, el tercer sujeto es no fumador de 54 años con tensión arterial sistólica de 140 mm/Hg, quien durante el periodo de observación no desarrolló enfermedad coronaria. Con el procedimiento *LOGISTIC* del paquete SAS, teniendo como entrada los datos de la tabla 5.3, se obtienen las estimaciones de β_0, β_1, β_2 y β_3, como se muestra en el siguiente modelo ajustado:

$$P(Y=1) = \frac{1}{1+\exp[20.1173 - 0.1496X_1 - 1.9477X_2 - 0.0770X_3]}$$

De manera que para una persona de 55 años, fumador y cuya tensión arterial sistólica es 150 mm/Hg, la probabilidad de padecer una enfermedad coronaria antes de terminar el periodo de estudio se estima de la manera siguiente:

$$\begin{aligned} P(Y=1) &= \frac{1}{1+\exp[20.1173 - 0.1496X_1 - 1.9477X_2 - 0.0770X_3]} \\ &= \frac{1}{1+\exp[20.1173 - 0.1496(55) - 1.9477(1) - 0.077(155)]} \\ &= 0.8802 \end{aligned}$$

Con este resultado se estima que, aproximadamente, al 86% de los sujetos que posean el perfil anterior se le presentará una enfermedad coronaria durante los 10 años de observación.

[3]Los datos han sido simulados.

Tabla 5.3: Enfermedades coronarias frente a tabaquismo, edad y TAS.

Sujeto	E. cor.	Edad	Fuma	TAS	Sujeto	E. cor.	Edad	Fuma	TAS
1	0	48	1	125	31	1	62	1	110
2	1	55	1	160	32	1	50	0	125
3	0	54	0	140	33	0	48	0	150
4	1	51	1	135	34	1	56	1	150
5	0	51	0	135	35	0	61	0	155
6	1	59	0	150	36	0	52	1	125
7	1	60	1	160	37	0	42	1	135
8	1	50	0	140	38	0	44	1	140
9	1	58	1	160	39	1	47	0	145
10	0	41	1	135	40	1	56	1	150
11	0	51	0	135	41	1	51	1	155
12	1	59	0	150	42	1	52	1	125
13	1	50	1	160	43	0	42	1	135
14	1	50	1	150	44	0	65	0	145
15	1	62	1	150	45	0	49	1	155
16	0	60	0	140	46	0	41	1	145
17	0	55	0	140	47	0	42	1	135
18	1	48	1	150	48	0	46	0	140
19	1	50	1	160	49	0	54	0	135
20	0	48	0	130	50	0	51	1	135
21	1	56	1	155	51	0	51	0	135
22	1	61	1	155	52	1	59	0	150
23	0	52	0	135	53	0	60	0	145
24	0	42	1	135	54	1	50	1	160
25	0	44	1	140	55	1	58	1	160
26	1	47	1	165	56	0	41	1	135
27	0	46	1	130	57	0	52	0	155
28	0	61	0	150	58	1	43	0	160
29	0	43	0	125	59	1	58	1	155
30	0	41	1	150	60	1	52	1	160

Puesto que $\exp(\beta_i)$ tiene la interpretación independiente mostrada en la sección (5.2), un intervalo de confianza para este parámetro puede ayudar a la interpretación. Este intervalo se elabora construyendo primero el intervalo de confianza para β_i, y aplicando luego la función exponencial a los extremos de este. El cómputo de los extremos para tal intervalo se muestra en la siguiente expresión:

$$\exp[\widehat{\beta}_i - 1.96\widehat{\sigma}(\widehat{\beta}_i)], \text{ y } \exp[\widehat{\beta}_i + 1.96\widehat{\sigma}(\widehat{\beta}_i)] \tag{5.10}$$

donde $\widehat{\sigma}(\widehat{\beta}_i)$ es el *error estándar* asociado al estimador $\widehat{\beta}_i$. El intervalo de confianza se interpreta de acuerdo con si este contiene o no el número 1. Si contiene a 1, se puede concluir que es igualmente riesgoso estar en un nivel de la variable que en el otro; es decir, la variable no tiene efecto significativo sobre la respuesta. Si el intervalo no contiene el 1, habrá que observar si los dos extremos de este son mayores o menores

que 1: si son mayores que 1, se concluye que tiene más riesgo el evento para el perfil de individuo dado por el numerador que el perfil dado por el denominador; si son menores que 1 se tiene la situación inversa.

Si se quiere saber qué influencia tiene el fumar o no, manteniendo constantes la edad y el TAS, en el desarrollo de enfermedades coronarias, por (5.8) se tiene:

$$\frac{O(X^k)}{O(X^l)} = \exp(\widehat{\beta}_2) = \exp(1.9477) = 7.012$$

Esto significa que una persona fumadora tiene 7 veces el riesgo de padecer enfermedades coronarias con relación a una persona no fumadora, con la misma edad y TAS de la anterior.

Una vez ajustado un modelo, la inquietud que suele aparecer es sobre la influencia de las diferentes variables en el resultado. Así, puede tenerse la inquietud de cuál de las p-variables influye más en el valor que tome $P(Y = 1)$. La alternativa de considerar importante aquella variable que tenga el valor más alto de $\widehat{\beta}_i$ es incorrecta, pues los valores de los β están asociados con las unidades de medición. Por ejemplo, para una variable cuyas unidades están en kilogramos, el coeficiente para la misma variable expresada ahora en gramos, es 1.000 veces el anterior. La manera usual y adecuada de cuantificar el aporte con fines comparativos es calcular, para cada variable, el producto entre la estimación de β_i y la raíz cuadrada de la varianza muestral de la respectiva variable X_i. De esta manera, el coeficiente que resulta es

$$\widehat{\beta}_i^* = \widehat{\beta}_i \widehat{\sigma}_i \tag{5.10}$$

el cual se conoce como *coeficiente estandarizado* para la variable X_i.

En resumen, los coeficientes estandarizados se recomiendan para comparar la magnitud de los efectos debidos a las diferentes variables sobre la probabilidad que se estima. Otra estadística es la razón de *odds*, que como se presentó anteriormente, corresponde a la exponencial calculada en el respectivo coeficiente de regresión. De manera que al ordenar de mayor a menor las razones de *odds* es equivalente a hacerlo con $\exp(\widehat{\beta}_1)$, $\exp(\widehat{\beta}_2)$, ..., $\exp(\widehat{\beta}_p)$, y a su vez equivale a ordenar directamente los $\widehat{\beta}_i$.

Para el caso de variables dicotómicas X_i, se sabe que $\widehat{\sigma}_i^2 = p_i(1 - p_i)$, donde p_i es la proporción de sujetos para los cuales $X_i = 1$. En este caso, el coeficiente estandarizado es igual a:

$$\widehat{\beta}_i^* = \widehat{\beta}_i \sqrt{p_i(1 - p_i)} \tag{5.10a}$$

Cuando todas las variables explicativas son de tipo dicotómico no se tiene la objeción hecha con relación a las unidades de medición. No obstante, para estas variables el orden relativo de los $\widehat{\beta}_i$ no se corresponde con el de los $\widehat{\beta}_i^*$, pues el coeficiente estandarizado de una variable dicotómica depende no solo de la magnitud de la estimación del coeficiente, sino de que el porcentaje de sujetos con el atributo se aproxime al 50% (pues la varianza es máxima cuando $p_i = 1 - p_i = 0.5$). La explicación de esto es lógica, ya que si los individuos, en un caso extremo, poseen el atributo, ha de ser nula la capacidad discriminatoria para distinguir entre los que van a tener el atributo y los que no; es decir, el parámetro no tiene fuerza explicativa (visto desde los datos muestrales).

El procedimiento LOGISTIC del SAS calcula estos coeficientes estandarizados, los cuales se notan por $\widehat{\beta}_i^{**}$, de manera diferente a los anteriores. La forma en que los calcula es la siguiente: divide la estimación del respectivo parámetro por la razón que resulta de dividir la desviación estándar de la distribución logística $(\pi/\sqrt{3})$ y la desviación estándar de la variable.

$$\widehat{\beta}_i^{**} = \widehat{\beta}_i/[\pi/\sqrt{3}/\widehat{\sigma}_i] = \frac{\sqrt{3}}{\pi}\widehat{\beta}_i\widehat{\sigma}_i = (0.5513)\widehat{\beta}_i^* \qquad (5.10b)$$

Note que en la expresión (5.10b) se muestra la relación entre los dos tipos de coeficientes estandarizados.

Las estadísticas básicas asociadas a las variables explicativas, para cada uno de los niveles de la variable respuesta (padecer o no enfermedad coronaria) y en forma global, se muestran a continuación:

Variable	E_Cor	Media	Desviación estándar
Edad	1	53.9286	5.1850
	0	49.3125	6.8888
	Global	51.4667	6.5290
Fuma	1	0.7500	0.4410
	0	0.4688	0.5070
	Global	0.6000	0.4940
Tas	1	150.2000	12.8000
	0	138.9000	8.3990
	Global	144.200	12.0100

Para los datos de la tabla 5.3 se ha desarrollado la estimación de los parámetros del modelo logístico, la razón de *odds* y los intervalos de confianza asociados a los parámetros. Estos resultados están contenidos en la tabla 5.4.

Tabla 5.4: Estimaciones máximo verosímiles con los datos de la tabla 5.3.

Variable	G. L.	Estimación	Error estándar	Estimador estandarizado	Razón *odds*	Intervalo de confianza	
(Intercept)	1	−20.11727	5.1663				
edad	1	0.14964	0.0564	0.5386	1.161	1.046	1.309
fuma	1	1.94768	0.7674	0.5305	7.012	1.713	36.720
tas	1	0.07699	0.0288	0.5097	1.080	1.023	1.149

De la tabla 5.4, por ejemplo, el coeficiente estandarizado asociado con la variable $FUMA$ se calcula de acuerdo con la igualdad (5.10b), así:

$$\widehat{\beta}_i^{**} = \frac{\sqrt{3}}{\pi}\widehat{\beta}_i\widehat{\sigma}_i = \frac{\sqrt{3}}{\pi}(1.9477)0.494032 = 0.5305$$

Ahora, como $\widehat{\beta}_i^{**} = (0.5513)\widehat{\beta}_i^{*}$, entonces:

$$\widehat{\beta}_i^{*} = \widehat{\beta}_i^{**}/(0.5513) = 0.9623$$

Los intervalos de confianza de Wald se calculan conforme a la expresión (5.9). Ninguno de los tres intervalos contiene a 1, en especial el intervalo asociado al parámetro $FUMA$; por tanto, se ratifica la afirmación sobre el riesgo de padecer enfermedades coronarias en los fumadores.

5.5 Variables ficticias (*dummy*)

Cuando se tienen variables explicativas de tipo categórico, el problema es cómo identificar las modalidades de las variables, de manera que se pueda cuantificar la influencia de estas sobre la probabilidad de ocurrencia de un evento. Variables como la raza, el tipo sanguíneo, la ocupación, resultan importantes al dar cuenta sobre la probabilidad de ocurrencia de un evento asociado a la salud.

El primer intento podría conducir a la asignación de números a las distintas modalidades o categorías. Por ejemplo, para el tipo sanguíneo una asignación podría ser:

Tipo de sangre	Valor de la variable
$\mathcal{A}$	1
$\mathcal{B}$	2
$\mathcal{AB}$	3
$\mathcal{O}$	4

Esta nominación no es apropiada, toda vez que la regresión logística interpretaría, por ejemplo, que tener sangre tipo O es cuatro veces tener sangre tipo $\mathcal{A}$ y el doble de tener sangre tipo $\mathcal{B}$, lo cual no tiene ninguna coherencia. A continuación se presenta una de las salidas más usuales a este problema. Otras alternativas se pueden consultar en Silva (1995, Cap. 5).

Suponga que la variable consta de k-modalidades. Se deben crear entonces $k-1$ variables dicotómicas, denominadas variables *dummy*,[4] las cuales se notan por $Z_1, Z_2, \ldots, Z_{k-1}$. A cada modalidad o categoría de la variable le corresponde un conjunto de valores de los Z_i con el cual se identifica dicha modalidad.

Una de las maneras más corrientes de asignar los valores por modalidad es la siguiente: si un individuo pertenece a la primera (aunque el término *primera* puede ser arbitrario) modalidad, la primera variable, Z_1, toma el valor 1 y las demás valen 0; para un individuo ubicado en la segunda modalidad, la variable Z_2 vale 1 y las demás 0; para la penúltima modalidad, la variable Z_k vale 1 y las demás 0. Para última modalidad, todas las variables Z_i valen 0.

La asignación de variables *dummy*, para el caso del grupo sanguíneo, queda como se muestra a continuación (se insiste que el orden puede ser arbitrario):

Grupo	Variables *dummy*		
sanguíneo	Z_1	Z_2	Z_3
$\mathcal{A}$	1	0	0
$\mathcal{B}$	0	1	0
$\mathcal{AB}$	0	0	1
$\mathcal{O}$	0	0	0

[4]La traducción más adecuada es variable *ficticia o artificial*.

La segunda alternativa es la siguiente:

Grupo	Variables *dummy*		
sanguíneo	Z_1	Z_2	Z_3
$\mathcal{A}$	0	0	0
$\mathcal{B}$	1	0	0
$\mathcal{AB}$	0	1	0
$\mathcal{O}$	0	0	1

La tercera alternativa define $k-1$ variables artificiales, pero con un esquema como el que se muestra a continuación:

Grupo	Variables *dummy*		
sanguíneo	Z_1	Z_2	Z_3
$\mathcal{O}$	0	0	0
$\mathcal{A}$	1	0	0
$\mathcal{B}$	1	1	0
$\mathcal{AB}$	1	1	1

Al ajustar un modelo de regresión logística que incluya una variable nominal que tiene k modalidades, esta se debe sustituir por $k-1$ variables *dummy*, y a cada una de ellas le corresponderá su respectivo coeficiente.

Como soporte para la interpretación de los coeficientes asociados a variables artificiales, considere un estudio[5] en el cual se evalúa el efecto del grupo sanguíneo en el posible padecimiento de cierta dolencia hematológica. Para esto se desarrolló una observación prospectiva en la que se incluyó el factor RH (codificado como 1 si es negativo y 0 si es positivo) como covariable a controlar. La tabla 5.5 contiene la información sobre 1.094 individuos. Para la variable grupo sanguíneo se han generado variables artificiales siguiendo el primer procedimiento descrito; a las modalidades de variable RH también se le asignaron variables ficticias. A continuación se muestra la sintaxis del paquete SAS con la cual: se introducen los datos, se definen las variables *dummy* y se desarrolla la estimación del modelo logístico para estos datos. Entre los símbolos $*/$ y $*/$ se expresa el significado de la sintaxis.

Por ejemplo, la modalidad grupo sanguíneo tipo $\mathcal{A}$ queda definida por $Z_1 = 1,\ Z_2 = 0\ y\ Z_3 = 0$. Las modalidades del factor RH se definen por $\mathrm{RH}_1 = 1$ para factor negativo y $\mathrm{RH}_1 = 0$ para factor positivo.

[5] Silva (1995, 35).

Tabla 5.5: Pacientes por grupo sanguíneo, RH y condición patológica.

	Respuesta			
	Enfermó (1)		No enfermó (0)	
	RH		RH	
Grupo sanguíneo	-(1)	+(0)	-(1)	+(0)
$\mathcal{O}$	50	60	26	48
$\mathcal{A}$	200	30	100	10
$\mathcal{B}$	150	60	75	19
$\mathcal{AB}$	100	64	52	50

```
DATA  Enferm;
INPUT tipo_s$ RH$ frecue Enfermo @@;
/*crea variable ficticia: sangre tipo A*/
Z_1=(tipo _s='A');
/*crea variable ficticia: sangre tipo B*/
Z_2=(tipo _s='B');
/*crea variable ficticia: sangre tipo AB*/
Z_3=(tipo _s='AB');
 /*crea variable ficticia: RH negativo */
RH_1=(RH='N');
CARDS;
O  N 50  1 O  P 60 1 O  N  26 0 O  P 48 0
A  N 200 1 A  P 30 1 A  N 100 0 A  P 10 0
B  N 150 1 B  P 60 1 B  N  75 0 B  P 19 0
AB N 100 1 AB P 64 1 AB N  52 0 AB P 50 0
;
/*invoca el procedimiento LOGISTIC*/
PROC LOGISTIC simple descending;
FREQ  frecue;  /*permite trabajar con los datos desde la
                 tabla de contingencia */
/*propone el modelo a estimar */
MODEL  Enfermo=Z_1  Z_2  Z_3  RH_1;
RUN;
```

Esta rutina suministra las estimaciones de los parámetros asociados al modelo propuesto, con los cuales se estima la probabilidad de padecer

la enfermedad, así:

$$P(Y=1) = \frac{1}{1+\exp[-0.3744-0.3161Z_1-0.3900Z_2-0.0700Z_3-0.0534RH_1]} \quad (5.11)$$

Para un individuo con RH negativo ($\text{RH}_1 = 1$) y sangre tipo B, por ejemplo ($Z_1 = 0,\ Z_2 = 1$ y $Z_3 = 0$), de la ecuación (5.11) se obtiene que

$$P(Y=1) = \frac{1}{1+\exp[-0.3744-0.3161(0)-0.3900(1)-0.0700(0)-0.0534(1)]} = 0.6938$$

Para un individuo con RH positivo ($\text{RH}_1 = 0$) y sangre tipo O ($Z_1 = 0,\ Z_2 = 0$ y $Z_3 = 0$), la probabilidad de padecer la enfermedad es

$$P(Y=1) = \frac{1}{1+\exp[-0.3744-0.3161(0)-0.3900(0)-0.0700(0)-0.0534(0)]} = 0.5925$$

La razón de *odds* para el factor RH, como se explicó anteriormente, según la ecuación (5.8), es igual a

$$\exp(\widehat{\beta}_4) = \exp(0.053) = 1.0544$$

lo cual significa que, para personas con el mismo grupo sanguíneo, quienes posean RH negativo tienen 5% más de riesgo hacia la enfermedad que las personas con RH positivo, independiente del grupo sanguíneo.

La interpretación de $\widehat{\beta}_2$ con relación al grupo sanguíneo, por ejemplo para el grupo B, es que $\exp(\widehat{\beta}_2)$ refleja la razón de *odds* correspondiente a tener sangre tipo B, respecto de los correspondientes a tener sangre tipo O, asumiendo que se ha controlado el factor RH. Para el caso

$$\exp(\widehat{\beta}_2) = \exp(0.390) = 1.4770$$

Así, tener sangre tipo B aumenta en 48% el riesgo de contraer la enfermedad con relación a tener sangre tipo O. En general, el exponencial de uno de los coeficientes de las variables *dummy* estima la magnitud en que aumenta (o disminuye) el riesgo de la ocurrencia del evento de interés, respecto de la modalidad que se haya tomado como *referencia* (la que vale cero en todas las variables *dummy*) cuando se mantienen constantes (bajo control) el resto de las covariables. Recuerde que este exponencial se compara con el número 1: valores por encima de 1 muestran que sujetos con la modalidad de interés tienen más riesgo que los individuos bajo la modalidad de referencia.

5.6 Ajuste del modelo

Una vez estimados los parámetros de un modelo, la inquietud se centra en la "importancia" de cada variable para el modelo y en el problema que se enfrenta. Algunos criterios que den cuenta de la bondad del modelo se revisarán a través de la verificación de la relevancia de cada variable en el modelo propuesto; algunas medidas de ajuste y la selección de las variables más relevantes. No obstante, se debe aclarar que el término relevante no necesariamente apunta a la importancia desde el punto de vista biológico o médico causal de una variable, sino a una visión estadística.

5.6.1 Contraste de hipótesis sobre los parámetros

Cuando se tiene un modelo con p variables y otro con $k < p$ variables, el problema es decidir cuál de los dos modelos se ajusta mejor a los datos. Al primer modelo se le nota por M y al más simple por M^*. La estadística de razón de verosimilitud es

$$G^2 = -2\ln\frac{L(M^*)}{L(M)} = -2\ln L(M^*) - 2\ln L(M) = G^2(M^*) - G^2(M) \tag{5.12}$$

La estadística G^2 mide los desvíos entre los datos (valores observados) y los valores ajustados por el modelo logístico, y se define en forma semejante a la que se muestra en la ecuación (2.17), es decir:

$$G^2 = 2\sum(\text{observ.})\ln\left(\frac{\text{observ.}}{\text{ajuste}}\right) \tag{5.12a}$$

La estadística dada en (5.12a) para comparar los dos modelos es simplemente la diferencia de los desvíos de estos dos modelos: la estadística es grande cuando el modelo M^* se ajusta poco con el modelo M.

Para muestras de tamaño grande, la estadística G^2 tiene distribución *ji–cuadrado*, con un número de grados de libertad igual a la diferencia entre los grados de libertad de los respectivos errores en los dos modelos, es decir, $gl = p - k$.

El cociente (o razón) de verosimilitud G^2 es útil para determinar si hay diferencia significativa entre incluir en el modelo todas las varia-

bles (*modelo saturado*) e incluir tan solo algunas de ellas. G^2 sirve para evaluar si las variables $X_1, X_2, \ldots, X_p$, integradas en conjunto al modelo, contribuyen más a "explicar" las modificaciones que se producen en $P(Y = 1)$ que con k de estas variables ($k < p$). La mayoría de los paquetes estadísticos muestran a G^2 descompuesto en la forma $-2 \ln L(M^*)$ y $-2 \ln L(M)$, donde $-2 \ln L(M^*)$ corresponde a la razón de verosimilitud del modelo ajustado únicamente por el intercepto.

Para el modelo relacionado con las enfermedades coronarias, el procedimiento LOGISTIC del SAS reporta los valores de la verosimilitud del modelo ajustado por el intercepto, $-2 \ln L(M^*) = 82.9$, y la verosimilitud del modelo saturado, $-2 \ln L(M) = 56.8$. La diferencia corresponde a $G^2 = 26.1$, la cual está asociada a una distribución ji–cuadrado de 3 grados de libertad (pues $p = 3$ y $k = 0$); el *valor p* es igual a 0.0002, con lo cual se puede afirmar que las variables $EDAD$, $FUMAR$ y TAS se ajustan adecuadamente al modelo.

En el mismo ejemplo, suponga que se han eliminado las variables $EDAD$ y $FUMAR$, de manera que $k = 1$. El mismo procedimiento LOGISTIC muestra los siguientes resultados para los dos modelos:

Modelo incompleto:	$-2 \ln L(M^*) = 68.0$
Modelo completo:	$-2 \ln L(M) = 56.8$

El cociente de verosimilitudes de los dos modelos es igual a

$$G^2 = 68.0 - 56.8 = 11.2$$

El percentil 95 de la distribución ji–cuadrado con $p - k = 3 - 1 = 2$ grados de libertad es 5.991, cantidad menor que G^2. Se concluye que las variables $EDAD$ y $FUMAR$ tienen información adicional para la variable explicativa superior a la que ofrece la variable TAS por sí sola.

Cuando la diferencia entre p y k es 1 ($p - k = 1$), se trata del caso en el que se verifica el aporte de una variable particular. Es decir, se quiere observar si la supresión de una variable específica reduce significativamente el grado de explicación que se obtiene cuando esta variable se incluye con las demás al modelo. Esto equivale a verificar la hipótesis $H_0 : \beta_i = 0$. La estadística con la cual se contrasta esta hipótesis es[6]:

$$= \frac{\widehat{\beta}_i}{\widehat{\sigma}(\widehat{\beta}_i)} \tag{5.13}$$

[6]Esta es la estadística de Wald (ecuación (4.17)).

donde $\widehat{\sigma}(\widehat{\beta}_i)$ es el *error estándar* de $\widehat{\beta}_i$. La estadística Z sigue una distribución normal estándar, de manera que se rechaza H_0, a un nivel de significancia del 5%, si $|Z| > 1.96$. El cuadrado de la estadística Z tiene distribución ji–cuadrado con un grado de libertad. Algunos paquetes estadísticos reportan el valor de Z^2 asociado a cada uno de los parámetros del modelo, junto con el respectivo *valor p*.

La tabla 5.6 muestra los valores de la estadística Z para verificar la significancia de cada uno de los parámetros del modelo sobre enfermedades coronarias (datos de la tabla 5.3). De acuerdo con los valores de la estadística de Z y los valores p respectivos, en un nivel de significancia $\alpha = 0.05$, todos los parámetros son significativos, en consecuencia, se muestra una asociación significativa de cada una de las variables $EDAD$, $FUMAR$ y TAS con el desarrollo de enfermedades coronarias durante el periodo de observación.

Tabla 5.6: Verificación de los parámetros, $H_0 : \beta_i = 0$.

Variable	G. L.	Estimación	Error estándar	Estadística Z	p-valor
Intercepto	1	−20.1173	5.1663	−3.89	0.0001
EDAD	1	0.1496	0.0564	2.66	0.0079
FUMA	1	1.9477	0.7674	2.54	0.0112
TAS	1	0.0770	0.0288	2.67	0.0076

5.6.2 Selección de modelos

En un modelo que incluya p variables, posiblemente no todas ellas sean relevantes para el problema. En tal caso se deben detectar las variables que menos aporten al modelo para decidir sobre su exclusión, y así obtener un modelo más simple (*parsimonioso*).

En el análisis de regresión existen varios procedimientos disponibles para la selección del modelo que, con el número más reducido posible de variables, se ajuste adecuadamente a los datos. La regresión *paso a paso*[7] es un procedimiento que consiste en construir sucesivos modelos de manera que cada uno difiera del precedente en una sola variable e ir comparando los resultados de cada versión con los de la anterior.

Existen dos versiones de la regresión paso a paso:

[7] En la literatura anglosajona se le denomina *Stepwise*.

a) La regresión "*hacia adelante*" (*forward*).

b) La regresión "*hacia atrás*" (*backward*).

Por la complejidad de los cálculos, estos procedimientos se explican para el caso de cuatro variables explicativas.

El primer procedimiento (hacia adelante) consta de los siguientes pasos:

Paso 1. Se ajustan cuatro modelos de regresión logística, uno con cada variable explicativa. Sean $G_1^2, G_2^2, G_3^2, G_4^2$ los cocientes de verosimilitud asociados a cada uno de los modelos (ecuación (5.12)).

Se identifica el mayor de estos cocientes. Suponga que fue G_2^2.

Paso 2. Se evalúa si G_2^2 es significativo, es decir, se compara con 3.84 (percentil 95 de la ji–cuadrado con un grado de libertad). Si es mayor que 3.84, X_2 se incorpora al modelo y se continúa con el paso 3. En caso contrario, se detiene el proceso sin selección alguna.

Paso 3. Se ajustan todos aquellos modelos con dos variables que contengan a X_2; para el caso hay tres posibles modelos: $\{X_1, X_2\}$, $\{X_2, X_3\}$ y $\{X_2, X_4\}$.

Se calcula la verosimilitud $-2\ln L$ para cada uno de los modelos $-2\ln L$. Así:

$$
\begin{aligned}
&-2\ln L_{12} \text{ para } X_2\ X_1\\
&-2\ln L_{23} \text{ para } X_2\ X_3\\
&-2\ln L_{24} \text{ para } X_2\ X_4
\end{aligned}
$$

Se identifica la pareja de variables para la cual $-2\ln L$ es menor. Suponga que corresponde al par $\{X_2, X_4\}$.

Paso 4. Se evalúa si $-2\ln(L_{24}/L_2)$ es mayor que 3.84. Si es cierto, se incluye X_4 en el modelo como segunda variable explicativa y se procede al paso quinto. En caso contrario, se termina la selección y se propone el modelo únicamente con la variable X_4.

Paso 5. Se ajustan todos los modelos con tres variables que incluyan las variables X_2 y X_4. Los posibles modelos tendrán como variables explicativas a $\{X_1, X_2, X_4\}$ o las variables $\{X_3, X_2, X_4\}$. Se

calcula $-2\ln L$ para los dos casos y se toma el menor de ellos. Suponga que es $-2\ln L_{241}$.

Paso 6. Se evalúa si X_1 hace un aporte significativo al grado de explicación que ya dan X_2 y X_4. Esto ocurre siempre que se cumpla que

$$-2\ln\frac{L_{241}}{L_{24}} > 3.84$$

En tal caso X_1 se adiciona a X_2 y X_4 y se va al último paso. De lo contrario, sólo quedan seleccionadas estas dos últimas variables.

Paso 7. Se ajusta el modelo completo con las variables: X_1, X_2, X_3 y X_4. Se calcula

$$-2\ln\frac{L_{2413}}{L_{241}}$$

Si este número supera al percentil de ji–cuadrado que se ha venido trabajando, 3.84, las variables se incluirían en el modelo. De lo contrario sólo quedarían las variables iniciales $\{X_1, X_2, X_4\}$.

El otro método es la regresión hacia atrás (*backward*), similar al anterior. Se ajusta la regresión logística para las $p = 4$ variables y se van ajustando modelos de orden inferior hasta llegar a uno que no pueda "degradarse" sin pérdida de información significativa. Se omite, por lo repetitivo, este algoritmo.

Un tercer procedimiento alternativo es el llamado procedimiento de las *significaciones sucesivas*. De manera esquemática consiste en lo siguiente:

Paso 1. Se ajusta el modelo con *todas* las variables.

Paso 2. Se identifican aquellas variables cuyos coeficientes, desde la estadística de Wald, sean significativamente diferentes de 0. Si todas resultan significativas, se concluye el proceso. En caso contrario, se sigue con el paso tercero.

Paso 3. Se ajusta el modelo con aquellas variables que resultaron significativas en el paso anterior y se vuelve al paso segundo.

Se debe definir lo que se entiende por *variables relevantes*, pues si el problema es hallar un modelo que optimice la predicción, la solución a

este problema ya ha sido abordada en esta sección; pero si el modelo se ajusta dentro de un marco conceptual (como debe ser), el proceso de selección de variables atañe más al área del conocimiento que se estudia que a la estadística. Esta anotación es oportuna, toda vez que muchos investigadores actúan como si los métodos estadísticos les eximieran de la experticia y conocimientos acerca del tema que tratan. Aquí "caen" bien las palabras de Tukey: *"estadístico es el que piensa con la cabeza de otro"*.

A continuación se muestra el procedimiento de selección de variables mediante el siguiente ejemplo extraído de SAS/STAT (1998). Los datos corresponden a la remisión o no de pacientes con cáncer a los cuales se les ha registrado una serie de variables explicativas. Como la intención es explicar, en forma general, el procedimiento de cómputo para seleccionar las variables, solo se presenta el nombre usado de las variables en el programa del procedimiento LOGISTIC. La rutina SAS con la que se procesa esta información se muestra al final de esta sección.

A continuación se transcribe la parte que corresponde a la selección de variables; se omite la escritura de cada uno de los pasos seguidos por el programa. La tabla 5.7 muestra en cada paso la variable que entra o es removida del modelo, la estadística ji–cuadrado, y el respectivo valor p bajo el cual entra o se remueve una variable.

Tabla 5.7: Summary of Stepwise Procedure

Step	Variable		Number	Score	Pr >
	Entered	Removed	In	Chi-Square	Chi-Square
1	LI1		1	7.9311	0.0049
2	TEMP		2	1.2591	0.2618
3	CELL		3	1.4701	0.2253

Con el nivel de significancia de 0.3, para que una variable entre o permanezca en el modelo se tiene que las variables LI, $TEMP$ y $CELL$ son las que finalmente conforman el modelo con el cual se predice la probabilidad de remisión por cáncer.

Para la selección tipo regresión hacia atrás (*backward*), la tabla 5.8 contiene el resumen del procedimiento (significancia de 0.2).

```
DATA remis;
INPUT remis cell smear infil li blast temp @@;
```

Tabla 5.8: Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Score Chi-Square	Pr > Chi-Square
1	BLAST	4	0.000844	0.9768
1	SMEAR	3	0.0951	0.7578
1	CELL	2	1.51135	0.2186
1	TEMP	1	0.6535	0.4189

```
CARDS;
1 0.8 0.83 0.66 1.9  1.1 0.996
1 0.9 0.36 0.32 1.4 0.74 0.992
0 0.8 0.88 0.7 0.8 0.176 0.982
0 1   0.87 0.87 0.7 1.053 0.986
1 0.9 0.75 0.68 1.3 0.519 0.98
0 1   0.65 0.65 0.6 0.519 0.982
1 0.95 0.97 0.92 1 1.23 0.992
0 0.95 0.87 0.83 1.9 1.354 1.02
0 1 0.45 0.45 0.8 0.322 0.999
0 0.95 0.36 0.34 0.5 0 1.038
0 0.85 0.39 0.33 0.7 0.279 0.988
0 0.7 0.76 0.53 1.2 0.146 0.982
0 0.8 0.46 0.37 0.4 0.38 1.006
0 0.2 0.39 0.08 0.8 0.114 0.99
0 1 0.9 0.9 1.1 1.037 0.99
1 1 0.84 0.84 1.9 2.064 1.02
0 0.65 0.42 0.27 0.5 0.114 1.014
0 1 0.75 0.75 1 1.322 1.004
0 0.5 0.44 0.22 0.6 0.114 0.99
1 1 0.63 0.63 1.1 1.072 0.986
0 1 0.33 0.33 0.4 0.176 1.01
0 0.9 0.93 0.84 0.6 1.591 1.02
1 1 0.58 0.58 1 0.531 1.002
0 0.95 0.32 0.3 1.6 0.886 0.988
1 1 0.6 0.6 1.7 0.964 0.99
1 1 0.69 0.69 0.9 0.398 0.986
0 1 0.73 0.73 0.7 0.398 0.986
;
RUN;
```

```
Title 'Regresión vía stepwise datos de remisión de
cáncer';
PROC LOGISTIC;
MODEL remis=cell smear infil li blast temp
/selection = stepwise slentry = 0.3
*/significancia para que una variable ingrese
(con forward y stepwise)*/
slstay = 0.3
*/significancia para que una variable permanezca
(con backward y stepwise)*/
details; */imprime cada paso del proceso*/
OUTPUT OUT=pred p=phat lower=lcl upper=ucl;
*/crea el archivo ‘‘pred’’*/
RUN;
Title 'Eliminación Backward';
MODEL remis= temp cell li smear
blast /selection=backward
slstay = 0.2 ctable;
RUN;
```

Se debe tener cuidado en el momento de interpretar estos resultados, pues se advierte que la técnica no distingue entre asociaciones de índole causal y las debidas a terceros factores involucrados en el proceso, incluso las que se observan como consecuencia de un sesgo en el estudio. Además, el proceso de selección se ampara en las pruebas de significación; por tanto, están sujetas a las suspicacias que despiertan. Una muestra de tamaño grande, el método, puede hacer que una variable quede incluida en el modelo, aunque no tenga mucha importancia biológica o clínica.

5.6.3 Bondad de ajuste

Cuando se ajusta un modelo estadístico a un conjunto de datos, el investigador debe preguntarse qué tan bien se ajusta el modelo propuesto a los datos, de manera que las conclusiones derivadas de este tengan, en principio, suficiente respaldo estadístico.

Suponga que se ha ajustado un modelo de regresión logística con n-sujetos. Sea p el valor que asume la función ajustada para un sujeto

determinado. Una vez estimado el modelo, se puede calcular este valor para cada uno de lo individuos incluidos en el estudio. Así, para cada individuo se dispone tanto de la probabilidad p que le ocurra el evento en cuestión como del valor Y (el verdadero resultado que para él se produjo). La evaluación del ajuste se desarrolla mediante las n-parejas (Y, p).

De manera intuitiva, el procedimiento para evaluar la bondad de ajuste mediante estos pares se describe a continuación: si el ajuste es adecuado, se espera que, con una alta frecuencia a valores grandes de p (cercanos a 1.0 se correspondan con $Y = 1$ y, recíprocamente, valores pequeños de p (cercanos a 0.0), se correspondan con $Y = 0$. Suponga que en una muestra de 400 individuos se obtienen los siguientes resultados:

para 100 individuos se obtuvo $p = 0.1$
para 100 individuos se obtuvo $p = 0.5$
para 200 individuos se obtuvo $p = 0.8$

Sobre estos datos, la pregunta es: ¿cuántos individuos con el resultado $Y = 1$ se esperan en cada uno de estos tres grupos si el ajuste es bueno? Si p es la probabilidad de que el evento ocurra, entonces el número de individuos que se espera "caigan" en cada uno de los tres grupos es 10, 50 y 160, respectivamente. De manera que si los números verdaderos de individuos ubicados en cada uno de estos grupos fuesen 9, 45 y 166, estas frecuencias observadas no diferirían mucho de las esperadas; es decir, se observaría una alta *concordancia* entre los resultados esperados con la regresión logística y los observados. Si las frecuencias observadas fueran 26, 9 y 185, se admitiría que la regresión logística se ajusta muy pobremente a los datos.

A continuación se describen (amparados en Silva (1995, 64–66)) los pasos que se deben seguir para medir la concordancia entre las predicciones del modelo y los datos reales.

1. Calcular las probabilidades de que ocurra el evento $(Y = 1)$ para cada uno de los n individuos p_1, p_2,..., p_n a partir del modelo estimado.

2. Ordenar esos n valores p_i de menor a mayor.

3. Dividir los valores de p_i $(i = 1, 2, \ldots, n)$ en grupos. Hay dos formas para la conformación de grupos: a) Dividir el ordenamiento

en cuartiles, deciles u otra forma similar (en este texto se asume que se hace con deciles), b) Formar el primer grupo con todos los sujetos para los cuales p sea menor que 0.1; el segundo, aquellos cuyos valores están entre 0.1 y 0.2, y así sucesivamente.

Sean n_1, n_2,..., n_{10} las frecuencias respectivas.

4. Sumar los valores de p en cada uno de los grupos conformados. Sean E_1, E_2,..., E_{10} son los *valores esperados*. De manera que E_i es la suma de los n_i valores de p correspondientes al i-ésimo grupo.

5. Contar en cada grupo el número de individuos para los cuales $Y = 1$. Estos son los *valores observados*. Se denotan tales valores por O_1, O_2,..., O_{10}. Una vez que se obtienen estos pares de valores, se computa la estadística:

$$\chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} + \sum_{i=1}^{10} \frac{(O_i^* - E_i^*)^2}{E_i*} \tag{5.14}$$

donde $E_i^* = n_i - E_i$ y $O_i^* = n_i - O_i$

La estadística χ^2 se distribuye como ji–cuadrado con 8 grados de libertad. La distribución de esta estadística no siempre es así, pues a veces los E_i o E_i^* son nulos o menores de 5. Cuando se tiene tal situación, el estadístico (5.14) no se computa. La estrategia más simple, en general, consiste en cotejar valores esperados y observados mediante simple inspección y evaluar el grado de *concordancia* entre valores esperados y observados.

El procedimiento LOGISTIC reporta esta estadística. Para los datos de la tabla 5.3, el valor reportado es: $Concordant = 79,6\%$.

5.7 Regresión logística con respuesta politómica

En algunas ocasiones se tienen variables respuesta con más de dos modalidades (*politómica*). En esta sección se trata el caso de ajuste de modelos de regresión para variables respuesta de tipo politómico. Se aborda primero el caso de la regresión logística *nominal*, que se usa

cuando no hay un orden natural entre las categorías de la respuesta; en segundo lugar, se trata la regresión logística ordinal, que se aplica cuando se puede establecer un orden natural obvio en las categorías de las respuestas, que debe ser tomado en cuenta en la especificación del modelo.

5.7.1 Regresión logística nominal

El problema en su esencia es el mismo que se ha venido tratando en este capítulo; la diferencia está en que la variable Y puede tomar los valores $1, 2, \ldots, r$ $(r \geq 2)$; asociadas a dos o más categorías.

Si se escribe $A_1,\ A_2, \ldots, A_r$ las r-modalidades de respuesta, entonces la variable aleatoria Y toma el valor 1 si el sujeto incurre en el evento A_1, Y vale 2 si le sucede A_2, y así sucesivamente. Suponga que se tienen p-variables explicativas[8] y que se desea expresar la probabilidad de que Y tome cada uno de estos r valores en función de dichas variables.

Se deben construir $r-1$ funciones (tantas como posibles modalidades de respuesta menos 1). De manera que se deben estimar $r-1$ conjuntos de $p+1$ parámetros cada uno (un conjunto por modelo. Note la similitud (caso especial) a la situación en que la variable respuesta es dicotómica. En resumen, el problema es la estimación de la probabilidad de pertenencia a cada una de las r-modalidades de la variable Y para un individuo que tiene un perfil determinado por los valores que asuman las variables $X_1,\ X_2, \ldots, X_p$.

Con el ánimo de ilustrar, suponga que se tiene una variable respuesta con $r = 3$ modalidades (posibles respuestas) y $p = 4$ variables explicativas. Esto implica que se deben estimar dos conjuntos de 5 parámetros cada uno:

$$\begin{array}{ccccc} \alpha_1 & \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \alpha_2 & \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \end{array}$$

Así, la probabilidad de que un sujeto esté en la primera modalidad, se calcula a través de la expresión

$$P(Y = 1) = \frac{\exp[\boldsymbol{\beta}_1' X]}{1 + \exp[\boldsymbol{\beta}_1' X] + \exp[\boldsymbol{\beta}_2' X]} \tag{5.15}$$

[8] En cada caso se hará la referencia a p como el número de variables explicativas o como un valor de probabilidad.

Para calcular la probabilidad de que el sujeto se encuentre en la segunda modalidad se emplea la fórmula

$$P(Y=2) = \frac{\exp[\boldsymbol{\beta}_2' X]}{1 + \exp[\boldsymbol{\beta}_1' X] + \exp[\boldsymbol{\beta}_2' X]} \tag{5.16}$$

donde:

$$\exp[\boldsymbol{\beta}_1' X] = \exp[\alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + \beta_{14}X_4]$$
$$\exp[\boldsymbol{\beta}_2' X] = \exp[\alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + \beta_{24}X_4]$$

Finalmente, la probabilidad de que el sujeto se ubique en la modalidad A_3 $(Y = 3)$ se calcula mediante el complemento de los dos eventos anteriores, es decir

$$P(Y=3) = 1 - P(Y=1) - P(Y=2) \tag{5.17}$$

La generalización al caso de r-modalidades y p variables explicativas es inmediata. Así,

$$P(Y=k) = \frac{\exp[\boldsymbol{\beta}_k' X]}{1 + \sum_{i=1}^{r-1} \exp[\boldsymbol{\beta}_i' X]};\ \text{para } k = 1, 2, \cdots, r-1 \tag{5.18}$$

donde

$$\exp[\boldsymbol{\beta}_i' X] = \exp[\alpha_i + \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots + \beta_{ip}X_p]$$

Para fijar algunos conceptos sobre este tipo de modelos, considere los datos de la tabla 1.2 del capítulo 1, los cuales corresponden a un ensayo clínico con el que se quiere investigar el progreso o mejoría conseguido al final del ensayo de un tratamiento para la artritis reumatoidea. Tanto a hombres como a mujeres se les asignó una actividad (tratamiento) o un placebo. Los datos se transcriben en la tabla 5.9. La respuesta de progreso (o mejoría) tiene estas tres modalidades: *Ninguno* (A_1), *Regular* (A_2) y *Bueno* (A_3).

Al ajustar el modelo, tomando la categoría *ningún progreso* como categoría de referencia se tiene:

Progreso	Intercepto	Sexo (femenino)	Trat. (actividad)
Regular	-2.77	1.66	1.11
Bueno	-2.51	1.38	2.17

Tabla 5.9: Datos de artritis.

Sexo	Tratamiento	Progreso			Total
		Bueno	Regular	Ninguno	
Femenino	Actividad	16	5	6	27
Femenino	Placebo	6	7	19	32
Masculino	Actividad	5	2	7	14
Masculino	Placebo	1	0	10	11

Fuente: Stokes, Davis y Koch (1997: 218).

Así, la probabilidad de que una persona de sexo femenino sometida al tratamiento (actividad) pertenezca a la categoría *progreso regular* es

$$\begin{aligned} P(Y=2) =& \frac{\exp[-2.77+1.66+1.11]}{1+\exp[-2.77+1.66+1.11]+\exp[-2.51+1.38+2.17]} \\ =& 0.2075 \end{aligned}$$

La probabilidad de que la misma persona pertenezca a la categoría *progreso bueno* es

$$\begin{aligned} P(Y=3) =& \frac{\exp[-2.51+1.38+2.17]}{1+\exp[-2.77+1.66+1.11]+\exp[-2.51+1.38+2.17]} \\ =& 0.5844 \end{aligned}$$

y la probabilidad de que no tenga progreso alguno es $P(Y=1) = 1-0.2075-0.5844 = 0.2081$. La probabilidad de que una mujer que no recibió tratamiento (placebo) pertenezca a la categoría *progreso regular* es

$$P(Y=2) = \frac{\exp[-2.77+1.66]}{1+\exp[-2.77+1.66]+\exp[-2.51+1.38]} = 0.1994$$

y la probabilidad de que una mujer que recibió placebo pertenezca a la categoría *progreso bueno* es

$$P(Y=3) = \frac{\exp[-2.51+1.38]}{1+\exp[-2.77+1.66]+\exp[-2.51+1.38]} = 0.1955$$

mientras que la probabilidad de que no tenga progreso es $P(Y=1) = 1-0.1994-0.1955 = 0.6051$. La probabilidad de que un hombre que no

recibió tratamiento (placebo) pertenezca a la categoría *progreso regular* es

$$P(Y=2)=\frac{\exp[-2.77]}{1+\exp[-2.77]+\exp[-2.51]}=0.0548$$

La probabilidad de que un hombre que no recibió tratamiento (placebo) pertenezca a la categoría *progreso bueno* es

$$P(Y=3)=\frac{\exp[-2.51]}{1+\exp[-2.77]+\exp[-2.51]}=0.071$$

Por último, la probabilidad de que un hombre que no recibió tratamiento (placebo) pertenezca a la categoría *ningún progreso* es $P(Y=1)=1-0.0548-0.0710$. Las demás probabilidades se obtienen de manera análoga.

5.7.2 Regresión logística ordinal

En muchas situaciones, las categorías de la variable respuesta tienen alguna clase de ordenamiento. En estos casos no es apropiado el uso de la regresión logística nominal, porque se podría perder la capacidad de detectar la forma en que la variable respuesta está relacionada con las variables independientes. De acuerdo con Dobson (2002), los modelos que se usan comúnmente cuando se tienen categorías ordinales son el modelo *logit acumulativo*, el modelo de *categoría adyacente*, el modelo *logit* de *continuación de razón* y el modelo de *odds proporcionales*; este último, por ser el que se implementa en la mayoría de los programas de análisis estadísticos, se trata a continuación.

El modelo de *odds* proporcionales se basa en el supuesto que el efecto de las covariables $X_1,\ldots,X_p$ es igual para todas las categorías en la escala logarítmica, el modelo es

$$\log\frac{\pi_1+\cdots+\pi_j}{\pi_{j+1}+\cdots+\pi_r}=\beta_{0j}+\beta_1X_1+\cdots+\beta_pX_p \qquad (5.11)$$

note que el productor lineal tiene un intercepto que depende de la categoría j, pero las otras variables explicativas no dependen de j.

Como en la regresión logística nominal, la razón de *odds* asociada a un incremento de una unidad en la variable explicativa X_k es $\exp(\beta_k)$ con $k=1,\ldots,p$

En el caso de una respuesta ordinal con tres categorías y dos variables explicativas, de acuerdo con el modelo (5.11), se tiene:

$$\log \frac{\pi_1}{\pi_2 + \pi_3} = \beta_{01} + \beta_1 X_1 + \beta_2 X_2 \tag{5.12}$$

$$\log \frac{\pi_1 + \pi_2}{\pi_3} = \beta_{02} + \beta_1 X_1 + \beta_2 X_2 \tag{5.13}$$

A partir de las ecuaciones (5.12) y (5.13), teniendo en cuenta que $\pi_1 + \pi_2 + \pi_3 = 1$, se demuestra que

$$P(Y = 1) = \frac{1}{1 + \exp\{-(\beta_{01} + \beta_1 X_1 + \beta_2 X_2)\}}$$

$$P(Y = 3) = \frac{1}{1 + \exp\{\beta_{02} + \beta_1 X_1 + \beta_2 X_2\}}$$

el valor de $P(Y = 2)$ se obtiene por diferencia

$$P(Y = 2) = 1 - P(Y = 1) - P(Y = 3)$$

Para ilustrar se usan los datos de la tabla 5.9. Note que en este caso existe un ordenamiento natural en las categorías de la variable respuesta mejoría: ninguna<regular<buena. Para la estimación de los parámetros de los dos modelos que se deben ajustar se construyen variables *dummy*. En este caso se tomará, para la variable género, 1 si es femenino y 0 en caso contrario; la variable tratamiento toma el valor 1 si corresponde a actividad y 0 en el otro caso

Mediante el siguiente programa, del procedimiento LOGISTIC del SAS, se obtienen los estimadores para los dos modelos.

```
DATA  artritis;
INPUT  genero$  tratam$  mejoria$  frecue @@;
trata=(tratam='ACTIV');
sexo=(genero='F');
CARDS;
F  ACTIV  BUENO  16  F  ACTIV  REGU    5
F  ACTIV  NINGUN 6   F  PLACE  BUENO   6
F  PLACE  REGU   7   F  PLACE  NINGUN  19
M  ACTIV  BUENO  5   M  ACTIV  REGU    2
```

```
M  ACTIV  NINGUN 7   M  PLACE  BUENO   1
M  PLACE  REGU   0   M  PLACE  NINGUN  10
;
PROC  logistic  order=data;
FREQ  frecue;
MODEL  mejoria=trata  sexo;
RUN;
```

Los resultados de la estimación se muestran en la tabla 5.10. A partir de

Tabla 5.10: Verificación de los parámetros, $H_0 : \ \beta_i = 0$.

Variable	G. L.	Estimación	Error estándar	Estadística de Wald	p-valor
Intercepto1	1	-2.6672	0.5997	19.7809	0.0001
Intercepto2	1	-1.8128	0.5566	10.6072	0.0011
TRATA(Act)	1	1.7973	0.4728	14.4493	0.0001
SEXO(F)	1	1.3187	0.5292	6.2102	0.0127

estas estimaciones, si se tiene un paciente hombre que estuvo sometido al tratamiento, la probabilidad de que no tenga mejoría (recuperación *ninguno*) es

$$\begin{aligned} P(Y=1) &= \frac{1}{1+\exp[-1.8128+1.7973]} \\ &= 0.5039 \end{aligned}$$

La probabilidad de que un individuo con el mismo perfil anterior tenga un estado de mejoría "bueno" se calcula como sigue:

$$\begin{aligned} P(Y=3) &= \frac{1}{1+\exp[-(-2.6672+1.7973)]} \\ &= 0.2953 \end{aligned}$$

Finalmente, la probabilidad de que un individuo tenga un estado de recuperación regular es $P(Y=2) = 1 - P(Y=1) - P(Y=3) = 0.2008$.

Es decir, la probabilidad de no recuperación es mucho más grande y la de alcanzar una mejoría regular o mediana sería la más baja.

Un caso más general es el que contempla variables explicativas además de categóricas las de tipo cuantitativo como la EDAD, el TAS, entre otras. Para estos casos, las estimaciones son semejantes.

En general las otras estadísticas, como la asociada razón de *odds*, las de prueba de hipótesis y bondad de ajuste no son tan sustancialmente diferentes a las tratadas para el caso dicotómico. Un tratamiento más extenso se puede consultar en Hosmer & Lemeshov (1989) y en Agresti (1996).

5.8 Algunas aplicaciones de la regresión logística

Entre los propósitos del modelamiento estadístico se cuentan la descripción de información, la predicción y el análisis de un problema contextualizado en una situación real.

La regresión logística no es ajena a los propósitos anteriores, pues con su aplicación se pretende tener un dispositivo con el cual se pueda hacer una descripción de la realidad contenida en un conjunto de datos sin perder de vista que esta mirada es reducida o parcial. Los estudios que tienen como propósito la *predicción* son de tipo prospectivo; entre estos se cuentan los que tienen que ver con el vaticinio del desarrollo de enfermedades crónicas. La finalidad analítica de la regresión logística se enmarca en aquellos estudios que pretenden explicar por qué ocurre algo; para el caso de la salud se intenta descifrar el entramado causal de una enfermedad (etiología).

Aunque en estas notas no se hace referencia explícita a la metodología con la cual se consiguen los datos, las conclusiones que se puedan extraer de los datos están asociadas al procedimiento con el cual se obtuvieron estos; llámese diseño muestral, diseño bioclínico (en general experimental), registro de información, entre otros, de suerte que los resultados deben inscribirse y contemplar tales contextos[9].

5.8.1 Descripción

La regresión logística puede emplearse con propósitos descriptivos. Por ejemplo, la ocurrencia de casos nuevos de una enfermedad (*inciden-*

[9]Uno debe prestar atención a la práctica médica; no solo a las teorías plausibles, sino también a la experiencia combinada con la razón, escribió Hipócrates.

cia), el número de casos de una enfermedad existentes en una población (*prevalencia*). De manera que la probabilidad que se estima mediante la regresión logística puede asumirse como una tasa de incidencia o de prevalencia, la cual depende de las variables explicativas puestas en el modelo. En general, se puede admitir que el riesgo de padecer una dolencia crece (o decrece) en la medida que se modifican ciertas variables (edad, índice de obesidad, altitud del sitio de residencia, sexo, zona de residencia).

A manera de ilustración (para describir la descripción) se toma, en sus partes esenciales, un ejemplo desarrollado por Silva (1995, 100-114) sobre lactancia materna.

5.8.2 Patrón de lactancia materna (LM)

La Organización Mundial de la Salud define *lactancia materna* como: "La ingesta de leche que excluye el consumo de cualquier otro alimento que no sea el que se deriva del pecho materno". Recomienda que los niños sean así alimentados hasta el cuarto mes de vida y que, en lo posible, la LM se extienda durante el primer año de vida.

Para conocer el patrón de duración de la LM se construye un modelo que esquematice o describa la realidad del consumo. Tal modelo puede ser la *curva de prevalencia*, es decir, una función que muestre el porcentaje de niños que aún consumen leche materna para cada edad considerada dentro del primer año de vida.

Para acopiar la información necesaria, en el estudio de la duración de la LM, en una comunidad definida en un marco epidemiológico, por razones prácticas y de costos, se descarta la idea de hacer un estudio prospectivo en el cual se registra la edad a la cual los niños de una cohorte abandonan la LM. De igual manera, un estudio retrospectivo en el que se reconstruye cómo se comportó el fenómeno, no es recomendable por la escasa fiabilidad que se pueda dar a unos datos que proceden de la memoria de los respondientes. Lo anterior lleva a la selección de una muestra que represente tal comunidad de niños menores de un año. Suponga que se tiene una muestra 1.000 de estos niños, a quienes se les ha registrado las variables "lactancia en el momento de la encuesta" (variable Y) y "edad al momento de la encuesta, en días" (variable X);

es decir,

$Y = 1$	si el niño aún lacta al momento de la encuesta
$= 0$	si el niño ya no lo hace
$X =$	edad (en días) del niño en el momento de la pregunta

Los datos de esta muestra se consignan en una matriz de 1.000 filas por 3 columnas. Con las 5 observaciones siguientes se pretende dar una idea sobre la escritura de todos los datos:

Niño	Lacta	Edad
1	0	145
2	0	113
3	1	61
4	1	27
5	1	177

Por ejemplo, la primera fila corresponde a un niño encuestado a los 145 días del nacimiento y que, en ese momento, ya había abandonado el seno materno; la cuarta fila representa a un niño que a los 27 días se mantenía lactando de la madre.

Con estos datos se debe estimar la función

$$P(Y = 1) = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 X)\right]}$$

Las estimaciones resultantes son $\widehat{\beta}_0 = 0.746$ y $\widehat{\beta}_1 = -0.0176$. La función que permite estimar la probabilidad de que un niño tipo de esta comunidad aún esté lactando a cierta edad X (estimación de prevalencia de la lactancia a esa edad) viene dada por:

$$P(Y = 1) = \frac{1}{1 + \exp\left[(-0.746 + 0.0176X)\right]}$$

La función anterior es una estimación de la *curva de prevalencia* del consumo de leche materna en el primer año. Su forma refleja el modo en que decrece la tasa de LM con el aumento de la edad, como se muestra en la figura 5.2. En esta se observa que la probabilidad de que un niño de 60 días de edad consuma leche materna es 0.42 (42%), es decir, que

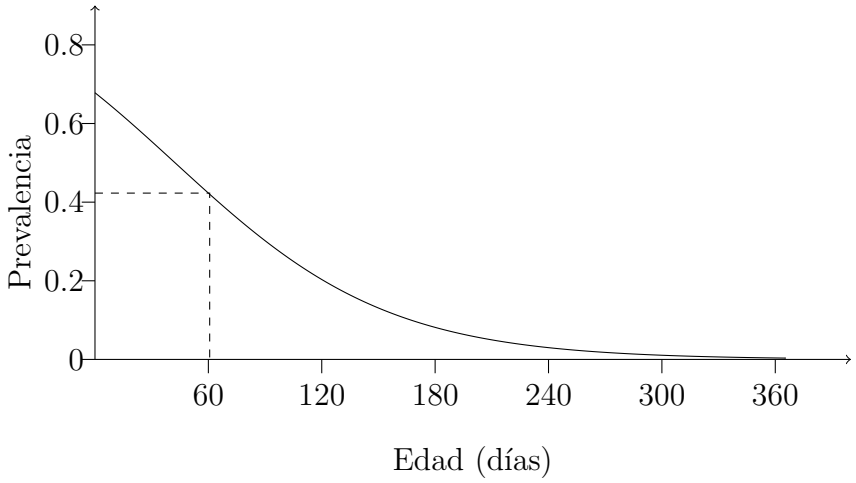

Figura 5.2: Curva de prevalencia de lactancia materna.

en esta población 42 de cada 100 niños[10], cuya edad sea 60 días, serán lactantes a esa edad.

El cálculo de $P(Y = 1)$ para diferentes valores de la variable edad señala la probabilidad de que un niño de determinada edad se encuentre lactando (prevalencia). La tabla 5.11 contiene las tasas de prevalencia para niños en edad 0 (recién nacido), 15, 30, 60, 90, 180, 240 y 365 días. En el

Tabla 5.11: Valores de la función logística.

Edad X	0	15	30	60
P(X=1)	0.68	0.62	0.55	0.49
Edad X	90	180	240	365
P(X=1)	0.30	0.08	0.03	0.003

caso de una regresión logística univariada, como el ejemplo que se trata aquí, el valor de X para el cual la probabilidad de poseer el atributo de interés sea $P(Y = 1) = 0.5$ corresponde a la edad *mediana* de abandono de la lactancia. A continuación se muestra el procedimiento para

[10] Aunque se ha considerado como base 100 niños, esta puede ser 1.000, 10.000 o cualquier otro valor, acorde con el respectivo tamaño de la población.

obtener el valor de X.

$$P(Y=1) = \frac{1}{1+\exp\left[-(\beta_0+\beta_1 X)\right]} = 0.5 \text{ (modelo logístico univariado)}$$

De manera que se debe encontrar el valor de X en la ecuación

$$\frac{1}{1+\exp\left[-(\beta_0+\beta_1 X)\right]} = 0.5,$$

que equivale a

$$\begin{aligned} 1+\exp\left[-(\beta_0+\beta_1 X)\right] &= 2 \\ \exp\left[-(\beta_0+\beta_1 X)\right] &= 1, \end{aligned}$$

Al aplicar logaritmos en ambos lados de la igualdad, se obtiene

$$-(\beta_0+\beta_1 X) = 0$$

de donde, al despejar X, resulta

$$X = -\left(\frac{\beta_0}{\beta_1}\right)$$

Esto quiere decir que un estimador para la mediana de abandono de la lactancia es $\widehat{X}_M = -\widehat{\beta}_0/\widehat{\beta}_1$. Para los datos disponibles, $\widehat{X}_M = 0.746/0.0176 = 42.4$, de donde se puede leer que 42.4 días es la edad mediana en que cesa el amamantamiento en la comunidad estudiada. De otra manera, que el 50% de la lactancia materna llega hasta los 42.4 días de edad, en consecuencia, que la otra mitad de los niños lactan 42.4 días o más[11].

Conforme al procedimiento anterior se pueden encontrar otros cuantiles como cuartiles, deciles o percentiles.

5.8.3 Comparación de curvas

En el caso que se quieran comparar los patrones de lactancia de niños con rasgos diferentes, como grupo socioeconómico de los padres, zona (rural o urbana), entre otras, se presentan a continuación tres alternativas para abordar esta situación.

[11] El diseño muestral que se siga permitirá hacer las respectivas inferencias.

Ajustes univariados independientes

El rasgo de interés divide la muestra en grupos (tantos como valores tome el rasgo). Para cada uno de estos se estima la regresión logística. La descripción se puede hacer mediante una gráfica de las curvas de regresión logística trazadas en el mismo sistema de ejes coordenados.

A manera de ejemplo, considere el caso en el que se comparan los patrones de lactancia correspondientes a dos modelos de atención primaria: el basado en el médico de familia (MF) y el que se basa en los policlínicos y médicos del sector (MS).

En Silva (1995, 106) se encuentra que, en una muestra de 1.401 niños, 823 eran atendidos por el sistema de médicos de familia, mientras que 578 estaban inscritos en un sistema de sectores de salud. La estimación de los respectivos parámetros y medianas se muestran en la tabla 5.12. La figura 5.3 muestra las curvas ajustadas con dos muestras independientes, por regresión logística, para la prevalencia de lactancia materna en los dos modelos de atención primaria.

Tabla 5.12: Estimación según régimen de atención primaria.

Régimen de atención primaria	*Estimación*		
	β_0	β_1	Mediana
Médico de sector (MS)	1.506	-0.0175	86.1
Médico de familia (MF)	1.441	-0.0128	112.6

Una conclusión que se puede derivar, por la observación de las dos curvas, es que los niños con médico de familia son más "demorados" en el abandono de la leche materna que los niños con médico de sector.

Ajuste multivariado

Este enfoque es más adecuado cuando las submuestras tienen tamaños relativamente pequeños. Se trata entonces de ajustar un modelo de regresión logística con dos variables explicativas: la variable edad y la variable que define la pertenencia a las subpoblaciones (tipo de atención). El modelo para $P(Y = 1)$ es

$$P(Y = 1) = \frac{1}{1 + \exp[-\beta_0 - \beta_1(edad) - \beta_2(AP)]}$$

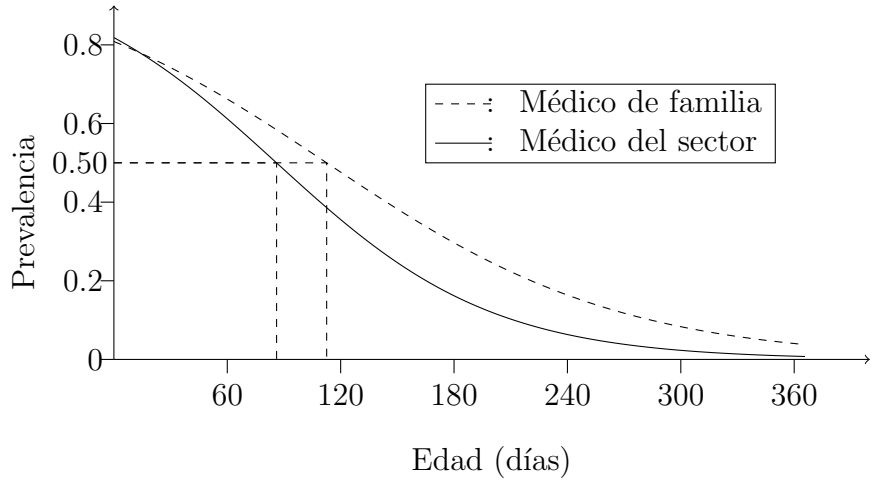

Figura 5.3: Curvas de prevalencia de lactancia materna por modelos de atención.

donde AP (atención primaria) vale 0 si el niño está bajo el modelo de MF, y 1 si está bajo el modelo de MS.

De esta forma, la expresión

$$\begin{aligned} P_{MS}(Y=1) &= \frac{1}{1+\exp[-\beta_0-\beta_1(edad)-\beta_2(1)]} \\ &= \frac{1}{1+\exp[-\beta_0-\beta_2-\beta_1(edad)]} \\ &= \frac{1}{1+\exp[-\beta_0^*-\beta_1(edad)]} \end{aligned}$$

con $\beta_0^* = \beta_0 + \beta_2$, es el patrón de lactancia bajo MS. Para $AP = 0$ se obtiene el patrón de lactancia bajo el sistema de atención por médico de familia; este es

$$\begin{aligned} P_{MF}(Y=1) &= \frac{1}{1+\exp[-\beta_0-\beta_1(edad)-\beta_2(0)]} \\ &= \frac{1}{1+\exp[-\beta_0-\beta_1(edad)]} \end{aligned}$$

Con los datos sobre los 1.401 niños se estimó el modelo con el siguiente

resultado:

$$P(Y=1) = \frac{1}{1+\exp[-(1.704 - 0.0164(edad) - 0.3020(AP))]}$$

Los modelos para la atención con MS y MF son, respectivamente,

$$P_{MS}(Y=1) = \frac{1}{1+\exp[-(1.704 - 0.3020 - 0.0164(edad))]}$$
$$= \frac{1}{1+\exp[-(1.402 - 0.0164(edad))]}$$

y

$$P_{MF}(Y=1) = \frac{1}{1+\exp[-(1.704 - 0.0164(edad))]}$$

La figura 5.4 indica que las curvas son "paralelas", es decir, que hay un retraso (durante el primer año de edad) del abandono de la lactancia materna del grupo MF sobre el grupo MS.

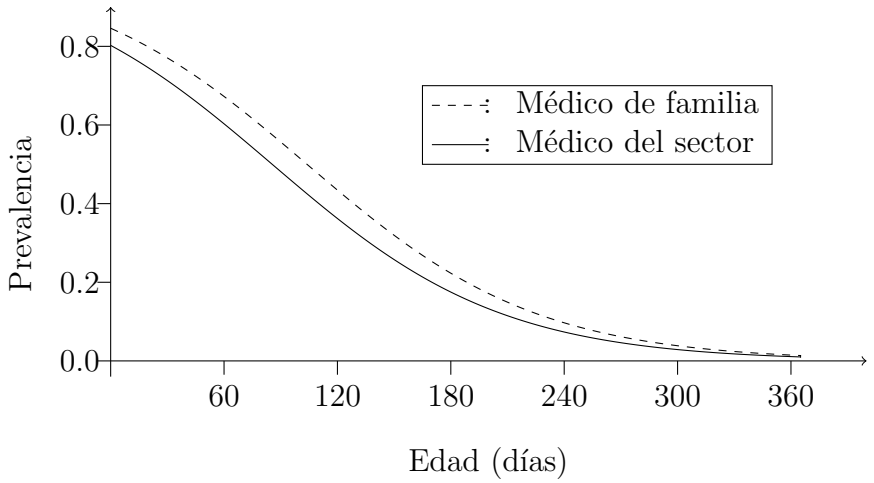

Figura 5.4: Curvas de prevalencia de lactancia materna. Ajuste bivariado

Las medianas, estimadas para cada grupo, son

$$\widehat{X}_{MF} = 103.9 \quad y \quad \widehat{X}_{MS} = 85.5$$

Ajuste con inclusión de interacciones en el modelo multivariado

Puede ocurrir que las diferencias entre las prevalencias de los dos modelos (MF y MS) sea diferente, tanto en signo como en magnitud, en el periodo considerado. Esto sugiere la presencia de un efecto conjunto de la edad y el grupo de atención primaria llamado *interacción*. En la gráfica esta situación se manifiesta con curvas logísticas que, además de diferir en la altura (abscisas), no tienen la misma velocidad con que estas diferencias se presentan respecto a iguales variaciones en el eje horizontal.

El modelo siguiente recoge esta posible característica

$$P(Y=1) = \frac{1}{1+\exp[-\beta_0 - \beta_1(edad) - \beta_2(AP) - \lambda(edad)*(AP)]}$$

En la expresión anterior $(edad)*(AP)$ representa la interacción entre la edad y el tipo de atención primaria. Note que si $AP = 0$, el modelo se reduce a

$$P_{MF}(Y=1) = \frac{1}{1+\exp[-\beta_0 - \beta_1(edad)]}$$

En cambio, si $AP = 1$, entonces la función logística es

$$\begin{aligned} P(Y=1) &= \frac{1}{1+\exp[-\beta_0 - \beta_1(edad) - \beta_2(1) - \lambda(edad)*(1)]} \\ &= \frac{1}{1+\exp[-(\beta_0+\beta_2) - (\beta_1+\lambda)(edad)]} \\ &= \frac{1}{1+\exp[-\beta_0^* - \beta^*(edad)]} \end{aligned}$$

con $\beta_0^* = \beta_0 + \beta_2$ y $\beta^* = \beta_1 + \lambda$.

Cuando no hay efecto conjunto (interacción) entre la edad y la atención primaria ($\lambda = 0$), coinciden las pendientes de las dos curvas, pero en presencia de interacción, las pendientes son diferentes.

Para los datos disponibles, el modelo ajustado es determinado por las siguientes estimaciones de los parámetros:

$$\widehat{\beta}_0 = 1.441, \quad \widehat{\beta}_1 = -0.0128, \quad \widehat{\beta}_2 = 0.0650 \quad y \quad \widehat{\lambda} = -0.0047$$

De aquí se obtiene que

$$\widehat{\beta}_0^* = 1.441 + 0.0650 = 1.506 \quad y \quad \beta^* = -0.0175$$

Al comparar estas estimaciones con las que se consignan en la tabla 5.12, se observa que el modelo con interacción desarrollado para estos 1.401 niños arroja curvas iguales para cada una de las modalidades de atención primaria, a las obtenidas independientemente con las dos submuestras de 823 y 578 niños, respectivamente.

No obstante, esto no siempre ocurre. Si las muestras son de tamaño grande, posiblemente las dos alternativas de modelos independientes y la de un modelo con interacción multivariado resulten equivalentes. En el caso que las muestras sean de tamaño pequeño, se sugiere el último procedimiento.

5.8.4 Índice de deserción

Hace referencia al porcentaje de niños que habiendo iniciado la LM al comienzo de su vida, la abandonan al cabo de cierto tiempo. Generalmente el interés se dirige a observar el *índice de deserción de la LM* en los primeros tres meses de vida.

Para el cálculo se debe tener la prevalencia en el momento del nacimiento y la prevalencia a los noventa días, es decir, $P(0)$ y $P(90)$. El indicador que refleja esta situación se denomina *índice de deserción* (ID) y se define por

$$ID = \frac{P(0) - P(90)}{P(0)}$$

$P(0) - P(90)$ es el porcentaje de niños que habiendo practicado LM desde su nacimiento, la han abandonado al cabo de 90 días. El ID expresa el porcentaje de los practicantes de LM que la abandonan a los tres meses. La tabla 5.13 muestra la prevalencia de LM al nacer y a

Tabla 5.13: Deserción de lactancia materna en los primeros tres meses para cuatro subpoblaciones.

Subpoblación	Prevalencia al nacer	Prevalencia a los 3 meses	Índice de deserción
Urbano - MF	58	21	64
Rural - MF	72	39	46
Urbano - MS	62	21	66
Rural - MS	70	31	56

los noventa días de vida; así como el índice de deserción ID para cuatro comunidades determinadas por la zona (urbana o rural) y el tipo de atención primaria (MF o MS).

Existen eventos que obligatoriamente deben ocurrir durante el desarrollo de determinado proceso, es decir, en condiciones normales, es ineludible el paso por tales eventos. El momento en el que ocurren estos sucesos se denomina *puntos singulares del desarrollo* (PSD). Algunos ejemplos de estos PSD son el abandono de la LM, la menarquia, la menopausia, el brote del primer molar permanente en un niño o la primera vez que consume carne. Cada uno de estos PSD determina un aspecto cualitativamente relevante en el desarrollo de un individuo.

Las curvas de prevalencia de individuos que han transitado por un PSD particular son importantes, tanto desde el punto de vista descriptivo como en la toma de decisiones en la práctica; es el caso del PSD asociado a la primera vez que un niño consiga caminar por sí mismo.

Las curvas de regresión logística son una herramienta útil para establecer normas que permitan decidir si es o no "atípico" que un individuo de cierta edad no haya pasado por el evento de interés. La calidad de las curvas construidas para estos propósitos está ligada a la calidad de la información recolectada, sea que se trate de una encuesta o de un diseño experimental.

A manera de ilustración suponga, en el ejemplo de la lactancia materna, que se recolectó información en la que, además de la edad y la condición de LM de cada niño, se indagó sobre el estado de cada niño con relación al consumo de los alimentos: zumos, verduras, carne y pescado. Para estimar el patrón que indique la introducción de un alimento por primera vez en la dieta de un niño, se construye la curva de prevalencia del primer consumo en función de la edad. Dado que el consumo crece cuando aumenta la edad, se esperan curvas de prevalencia con pendiente positiva (crecientes). Las curvas de regresión logística que representan la prevalencia del consumo de estos cuatro alimentos a lo largo del primer año se muestran en la figura 5.5.

Los organismos de salud dicen que un niño que adquiere hábitos nutricionales adecuados mediante el suministro de los alimentos en forma paulatina, optimizará su desarrollo en las respectivas fases de crecimiento. Para contrastar el cumplimiento con las normas internacionales con relación a la introducción de estos cuatro alimentos, se tienen los si-

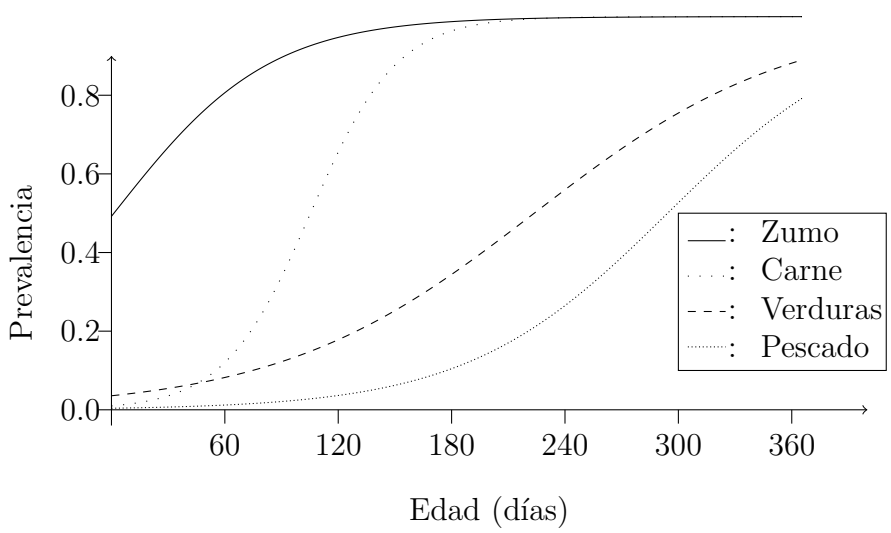

Figura 5.5: Curvas de prevalencia de consumo de cuatro alimentos.

guientes intervalos de edad:

Zumos	de 120 a 150 días
Verduras	de 150 a 180 días
Carne	de 180 a 210 días
Pescado	de 300 a 360 días

Si el evento "*consumo del alimento de interés por primera vez*" ocurre fuera de estos intervalos, se estaría incumpliendo la recomendación que al respecto hacen las autoridades de salud. El incumplimiento puede darse por la introducción precoz del alimento o por su introducción tardía. Para cuantificar el incumplimiento de estas recomendaciones, se estiman las tasas de prevalencia al comienzo y al final de estos intervalos de tiempo; el ideal es que estas prevalencias sean 0 y 1, respectivamente.

Suponga que un alimento debe introducirse entre los X_1 y X_2 días de nacido, y que P_1 y P_2 son las tasas de prevalencia de consumo a esas edades, respectivamente. P_1 es la probabilidad de que un niño haya iniciado su consumo del alimento antes de la edad X_1, y $1 - P_2$ es la probabilidad de que un niño no lo haya hecho a la edad X_2. La cantidad $P_1 - P_2$ representa el porcentaje de niños que reciben el alimento dentro del periodo establecido.

Las curvas de la figura 5.5 muestran el distanciamiento notorio de los ideales de ablactación[12] para los cuatro alimentos, puesto que predomina su introducción a edad temprana en los niños, con excepción de la verdura cuyo incumplimiento es por retraso.

5.8.5 Estudios prospectivos

Un estudio prospectivo es aquel en el que la exposición al factor de riesgo y los resultados de salud se observan después de comenzar el estudio. En estos estudios se inicia con un conjunto de individuos en un momento del proceso, cuyo desenlace de interés aún no se ha observado. Los individuos pueden ser separados en grupos de acuerdo con algunos rasgos, por ejemplo, naturales (mujeres gestantes fumadoras y mujeres gestantes no fumadoras) o condiciones introducidas o "manipuladas" por el investigador (tipos de medicamentos para tratar una dolencia).

El primer caso (rasgos naturales) es el estudio conocido como de *cohorte*, mientras que el segundo (tratamientos) corresponde a los estudios *experimentales*, como los ensayos clínicos.

5.8.6 Estudios de cohorte

Es un estudio en el que a un conjunto específico de individuos se le hace una observación de seguimiento.

Suponga que se han observado 2.000 personas infartadas que ingresan en un servicio de cuidados intensivos. Para cada paciente se registra la siguiente información:

Fuma: 1 si es fumador y 0 si no lo es.

Edad: 1 si es mayor de 64 años, 0 si tiene 64 años o menos.

Se evalúa si el individuo muere en el centro de salud (muerte=1) o si egresa vivo (muerte=0). Este es un caso de cohorte; los resultados se consignan en la tabla 5.14. Una pregunta que se debe responder se refiere a la asociación existente entre el tabaquismo y la muerte de estos pacientes. Una estrategia para la búsqueda de una respuesta a este

[12] Proceso de destete de los bebés.

Tabla 5.14: Cohorte de 2.000 pacientes infartados.

	Muerte=1		*Muerte=0*	
Edad	Fuma=1	Fuma=0	Fuma=1	Fuma=0
Edad=1	260	80	500	160
Edad=0	40	20	600	340

interrogante es el análisis mediante la estadística ji–cuadrado para tablas de contingencia (sección 2.3). Por ejemplo, para la asociación entre el tabaquismo y el deceso de estos pacientes, se encuentra que $\chi_0^2 = 5.95$, el cual, comparado con 3.84 (el percentil 95 con 1 grado de libertad de la distribución ji–cuadrado), muestra que los datos son evidencia suficiente para establecer la asociación entre estas variables. Por un procedimiento semejante se verifican asociaciones entre la edad y el deceso y entre los tres: tabaquismo, edad y deceso. De acuerdo con la tabla 5.15 y con la

Tabla 5.15: Decesos por tabaquismo.

Tabaquismo	Mueren	Sobreviven	Total
Fumaban (F)	300	1.100	1.400
No fumaban (NF)	100	500	600
Total	400	1.600	2.000

expresión (2.43), el riesgo relativo de muerte por tabaquismo es

$$RR = \frac{P_F(Muerte)}{P_{NF}(Muerte)} = \frac{300/1400}{100/600} = \frac{0.21}{0.16} = 1.29$$

Por el mismo procedimiento anterior, el riesgo relativo de muerte de

Tabla 5.16: Decesos por edad.

Edad	Mueren	Sobreviven	Total
Tercera edad (T)	340	660	1.000
65 o menos (NT)	60	940	1.000
Total	400	1.600	2.000

estos infartados atribuible a la edad es

$$RR = \frac{P_T(Muerte)}{P_{NT}(Muerte)} = \frac{340/1.000}{60/1.000} = \frac{0.34}{0.06} = 5.6$$

Este valor de RR es tan contundente que hace innecesaria la utilización de otra estadística para evidenciar la existencia de una asociación entre el deceso y la edad de las personas infartadas que conforman la cohorte de este estudio.

El análisis de la influencia de la variable tabaquismo y de la variable edad sobre el deceso se ha hecho de forma separada. La inquietud a resolver es si la diferencia entre las tasas de muerte para fumadores y para no fumadores es atribuible efectivamente al hecho de que el tabaquismo afecta apreciablemente el riesgo de muerte, o que su efecto real queda incrementado por la participación de la variable edad. Un análisis para esta situación sería como la que se presenta en el capítulo 2 para el caso de asociación entre tres variables.

Un análisis equivalente puede hacerse mediante la regresión logística para predecir la probabilidad de deceso o no en función del tabaquismo y la edad del paciente. De acuerdo con las variables explicativas, se pueden construir tres modelos: en función del tabaquismo, de la edad o de los dos (tabaquismo y edad). La tabla 5.17 exhibe las estimaciones para cada uno de los tres modelos. Para el modelo 1, de acuerdo con las

Tabla 5.17: Modelos ajustados para 2.000 infartados.

Modelo	Variables explicativas	Coeficientes	Error estándar
1	Tabaquismo	$\widehat{\beta}_0 = -1.609$	
		$\widehat{\beta}_1 = 0.310$	0.127
2	Edad	$\widehat{\beta}_0 = -2.751$	
		$\widehat{\beta}_1 = 2.088$	0.149
3		$\widehat{\beta}_0 = -2.79$	
	Tabaquismo	$\widehat{\beta}_1 = 0.060$	0.137
	Edad	$\widehat{\beta}_2 = 2.081$	0.150

expresiones (5.6), (5.7) y (5.8) $\exp\{\widehat{\beta}\}$ se interpreta como una estimación del riesgo relativo de muerte debido al hábito de fumar. El resultado es:

$$\exp\{\widehat{\beta}\} = \exp\{0.310\} = 1.36$$

cifra bastante cercana a 1.29, calculada mediante la tabla 5.15.

Para el modelo 2 se obtiene:

$$\exp\{\widehat{\beta}\} = \exp\{2.088\} = 8.07$$

El riesgo relativo calculado de la tabla 5.16 es 5.6, el cual es significativamente diferente al calculado para el modelo 2. Esto se debe a que, en este caso, las tasa de casos para los cuales ocurre la muerte ($Y = 1$) es muy alta (400 de los 2.000 infartados mueren).

Si para el modelo 3 se considera el coeficiente de la variable tabaquismo, el resultado es

$$\exp\{\widehat{\beta}_1\} = \exp\{\ 0.060\} = 1.07$$

cifra que se lee como la estimación del riesgo relativo de fumar al controlar el efecto de la edad. Este resultado permite afirmar que la asociación entre el hábito de fumar y el infarto mortal se "diluye" cuando se controla la edad.

5.8.7 Ensayos clínicos

Los ensayos clínicos son estudios en los que el investigador "modifica" o interviene la realidad, de acuerdo con un plan o diseño experimental, para observar algunos resultados que le interesan. Aunque la aplicación más usual está asociada a experimentos del ámbito biomédico, la metodología que aquí se esboza puede emplearse en otros campos experimentales como el social, el educativo, el psicológico e incluso en ingeniería.

La aplicación más corriente es aquella en la que se quiere evaluar el efecto de ciertos tratamientos terapéuticos, como medicamento, dieta, dosis de un fármaco, procedimiento quirúrgico, etc. La efectividad de tales tratamientos se hace mediante la comparación con tratamientos testigo, como un placebo, un método cultural o convencional, etc. Además de considerar la asignación aleatoria de sujetos experimentales a cada uno de los tratamientos y a otros requerimientos de tipo experimental (como los bloques), es importante definir la *variable respuesta* a considerar. Para efectos de emplear la regresión logística, se asume la variable como dicotómica, es decir la recuperación (éxito: $Y = 1$) o no (fracaso: $Y = 0$) del individuo.

Uno de los procedimientos seguidos para verificar la efectividad de un tratamiento es la comparación de las tasas de "éxito" (por ejemplo,

mejoría o alivio) mostradas por cada tratamiento. El problema en estos casos es garantizar que las diferencias se deben principalmente a los tratamientos y no a otros factores no explicitados que favorezcan el resultado de algún tratamiento. Las estrategias básicas del diseño experimental, como: *asignación aleatoria* de los tratamientos a las unidades experimentales, el *control local* de alguna(a) variables(s) (bloqueo u homogeneización) y la *replicación* de los tratamientos permiten asegurar que las respuestas observadas sobre los distintos grupos se distinguen solamente por el hecho de haber aplicado en cada uno de ellos tratamientos específicos.

En las condiciones experimentales señaladas anteriormente, se puede desarrollar la verificación de hipótesis con relación al efecto de cada uno de los tratamientos (a la manera de un análisis de varianza). Pero no siempre se dispone del suficiente número de unidades experimentales para hacer válidas las comparaciones; puede darse que la asignación de los tratamientos no haya sido realmente aleatoria. Aún más, con los métodos tradicionales para el análisis de estos datos consideraría el análisis de tablas de contingencia (capítulo 2), los cuales, mediante la estadística ji–cuadrado o afines, tratarían de revelar la posible asociación entre la mejoría y las variables explicativas en consideración. En estas circunstancias, la regresión logística es una alternativa para el análisis de este tipo de datos.

Suponga que se quiere evaluar la capacidad analgésica de la acupuntura (Silva 1995, 150-154) en pacientes que padecen dolores lumbares. Para descartar que la eventual mejoría reportada por el paciente sea de tipo sugestivo (efecto placebo), se establecen dos grupos: uno en el que se aplica tratamiento formal con acupuntura y otro al que también se le trata con agujas, pero no aplicadas en los "puntos teóricos" establecidos para esta terapia, sino en otros ubicados aleatoriamente.

La respuesta es $Y = 1$ si después de dos meses de tratamiento el individuo declara que experimenta una notable mejoría, y $Y = 0$ en caso que manifieste que los dolores se mantienen o aumentan.

Se consideran las siguientes variables del modelo: el tratamiento (X_1), la cual vale 1 si el individuo se somete al procedimiento real y 0 si se le aplicó el procedimiento falso; el sexo (X_2); la edad (X_3) y la historia de consumo de psicofármacos (X_4). A continuación se muestra un cuadro resumen de estas variables:

Y 1 si se presenta mejoría, 0 en caso contrario
X_1 1 si el individuo es sometido al tratamiento, 0 en caso contrario
X_2 1 si es mujer, 0 si es hombre
X_3 edad en años
X_4 consumo de psicofármacos con las siguientes categorías:
0: ningún consumo
1: consumo ocasional
2: consumo regular
3: uso frecuente
4: fármaco dependiente

En la tabla 5.18 se muestra la estructura de los datos asociados a las variables anteriores, medidas sobre 80 individuos (40 del grupo tratamiento verdadero y 40 del grupo placebo). Por ejemplo, el primer individuo es un hombre de 49 años, consumidor ocasional de fármacos a quien se le aplicó acupuntura de manera simulada (placebo) y manifiesta tener una mejoría; mientras que el individuo 80 es un hombre de 33 años, consumidor ocasional de fármacos a quien se le aplicó acupuntura verdadera y manifiesta no tener mejoría. Con la regresión logística se hace

Tabla 5.18: Esquema de datos sobre un ensayo clínico de acupuntura.

Sujeto	Tratamiento (X_1)	Sexo (X_2)	Edad (X_3)	CFármacos (X_4)	Respuesta (Y)
1	0	0	49	1	1
2	0	1	35	0	0
3	0	1	23	0	0
⋮	⋮	⋮	⋮	⋮	⋮
79	1	0	43	4	1
80	1	0	33	1	0

Fuente: Silva (1995, 152)

un ajuste de los datos en el que la probabilidad de respuesta se pone en función de las variables: tratamiento (X_1), sexo (X_2), edad (X_3) y consumo de psicofármacos (X_4). La tabla 5.19 contiene los resultados de las estimaciones de los parámetros junto con su error estándar y el valor de la estadística de Wald (ecuación 5.13). La estadística de Wald para cada una de las tres primeras variables muestra que los respectivos

Tabla 5.19: Ajuste de la probabilidad de mejoría.

Variables	Estimación	Error est.	Z de Wald
Intercepto	-4.658	-	-
Tratamiento	0.778	0.807	0.964
Sexo	1.452	0.934	1.554
Edad	-0.019	0.033	-0.567
Cons. farm.	1.913	0.460	4.155

parámetros asociados no son significativamente diferentes de cero (el Z Wald es menor en valor absoluto que 1.96), una vez que se controlan las otras tres variables. No obstante, cuando se controlan las primeras tres variables, la variable X_4 muestra una relación positiva y altamente significativa con la mejoría ($Z = 4.155$). Una lectura de este resultado es la siguiente: independientemente del tratamiento que reciban (real o placebo), aquellos individuos con mayor historia de consumo de psicofármacos manifiestan mejoría con más frecuencia.

De la tabla 5.19 se pueden calcular algunas razones de *odds*. Por ejemplo, para comparar el caso en que se aplica el tratamiento real ($X_1 = 1$) en mujeres ($X_2 = 1$) que consumen ocasionalmente psicofármacos ($X_4 = 1$) con hombres de la misma edad que nunca consumen psicofármacos ($X_2 = 0$ y $X_4 = 0$) y se les aplica el mismo tratamiento, se calcula:

$$\exp[\widehat{\beta}_1 + \widehat{\beta}_2] = \exp[1.452 + 1.913] = \exp[3.365] = 28.9$$

lo cual significa que es casi treinta veces más probable que se obtengan resultados favorables en el primer caso que en el segundo. Esto último merece esta aclaración: los *odds* en el primer caso son casi treinta veces mayores que los del segundo, pero la diferencia entre la razón de *odds* y el riesgo relativo generalmente es mínima y los conceptos son bastante cercanos.

5.8.8 Estudios caso-control

Son estudios en los cuales a un grupo de individuos (en general, animales, plantas u objetos) que tienen una condición determinada (*casos*) son considerados para compararlos con un grupo de individuos que no tienen dicha condición (*controles*); generalmente la condición es una en-

fermedad, una malformación, una disfunción neurológica, etc. De aquí el nombre de estudios *caso-control* o *retrospectivos*[13].

Uno de los usos frecuentes en investigación epidemiológica es la regresión logística, con la cual se obtienen razones de *odds* ajustadas a partir de las estimaciones de los coeficientes de las pendientes (los β), para variables respuesta medidas sobre grupos caso-control.

Para cada uno de los grupos se registran atributos (por ejemplo, factores de exposición) considerados influyentes en el resultado de la respuesta que se investiga.

5.8.9 Razón de *odds* y riesgos relativos

Nuevamente, considere que $Y = 1$ corresponde a un sujeto que desarrolló la condición de interés para el estudio y $Y = 0$ si no la desarrolló. En los estudios retrospectivos se tiene $Y = 1$ para los casos y $Y = 0$ para los controles. La intención con la construcción de regresión logística con p variables independientes (regresoras) es buscar una explicación de la respuesta (Y) en las variables que se ponen en juego desde el marco conceptual asumido.

Salvo en circunstancias poco frecuentes, la muestra conformada por los casos y los controles incluidos en un estudio representa la población objeto de estudio. Como consecuencia, en estos casos la función logística no permite tener estimaciones de $P(Y = 1)$ para un perfil dado.

Una ilustración es el caso de una enfermedad que se presenta en una población con una prevalencia, por ejemplo, de 2 sujetos por cada 1.000. Un estudio prospectivo, por ejemplo, requeriría muchos miles de individuos para conseguir unos pocos casos de individuos con la dolencia. Una situación típica es aquella en la que se estudian 200 casos y 400 controles, pues mientras que en la población hay un individuo con la enfermedad por cada 499 que no la padecen, en la muestra hay 1 enfermo por cada 2 sujetos sanos.

La estimación de $P(Y = 1)$ mediante los $\widehat{\beta}$, requiere una corrección a la estimación de $\widehat{\beta}_0$ (Hosmer y Lemeshow, 1989: 180). La corrección que

[13]Estudios que utilizan la información sobre exposición y estados de salud previos.

se sugiere como la más adecuada es

$$\widehat{\beta}_0^* = \widehat{\beta}_0 - ln(\pi_1/\pi_2) \tag{5.19}$$

donde π_1 y π_2 son las fracciones de muestreo con que se eligieron los casos y los controles, respectivamente. Es fácil verificar que si n_1 y n_2 es el número de casos y controles, respectivamente, y si f es la tasa de prevalencia de la enfermedad estudiada, entonces

$$\frac{\pi_1}{\pi_2} = \frac{n_1}{n_2}\frac{1-f}{f} \tag{5.20}$$

El procedimiento consiste en computar la función logística empleando $\widehat{\beta}_0^*$ en lugar de $\widehat{\beta}_0$; así,

$$P(Y=1) = \frac{1}{1+\exp[-\widehat{\beta}_0^* - \widehat{\beta}_1 X_1 - \cdots - \widehat{\beta}_p X_p]}$$

En la situación señalada se tiene que de cada 1.000 individuos en la población hay 2 enfermos y 998 no enfermos: $f = 0.002$ y $1-f = 0.998$.

Entonces de (5.20) se obtiene:

$$\frac{\pi_1}{\pi_2} = \frac{200}{400}\frac{0.998}{0.0002} = \frac{998}{4} = 249.5$$

En consecuencia, de acuerdo con la corrección para $\widehat{\beta}_0$ anterior,

$$\widehat{\beta}_0^* = \widehat{\beta}_0 - ln(\pi_1/\pi_2) = \widehat{\beta}_0 + ln(249.5) = \widehat{\beta}_0 + 5.52$$

Para ilustrar el uso de la regresión logística en estos estudios, se retorna a los datos de la tabla 5.14. Asuma que otro investigador decide hacer un estudio caso-control con estos datos. Dentro del esquema de estos estudios, decide usar los 400 casos (es decir, todos infartados fallecidos) y seleccionar 400 controles entre los 1.600 infartados que no fallecieron. Suponga que los resultados de esta muestra son los contenidos en la tabla 5.20. El ajuste de la regresión logística a estos datos, para cada uno de los tres modelos considerados anteriormente, se muestra en la tabla 5.21.

La estimación de los coeficientes($\widehat{\beta}_1$ y $\widehat{\beta}_2$) de las variables es casi igual; así, $exp[\widehat{\beta}_i]$ es una estimación de la razón de los *odds* asociados a un individuo que toma cierto valor sobre los *odds* correspondientes a otro

Tabla 5.20: Resultados de un estudio caso-control para evaluar letalidad en infartados con hábito de fumar y edad como factores explicativos.

	Casos ($Y = 1$)		*Controles* ($Y = 0$)	
Edad	Fuma=1	Fuma=0	Fuma=1	Fuma=0
Edad=1	260	80	125	40
Edad=0	40	20	150	85

Tabla 5.21: Modelos ajustados para 400 casos y 400 controles.

Modelo	Variables explicativas	Coeficientes	Error estándar	Z *de Wald*
1	Tabaquismo	$\widehat{\beta}_0 = -0.223$		
		$\widehat{\beta}_1 = 0.310$	0.158	1.963
2	Edad	$\widehat{\beta}_0 = -1.365$		
		$\widehat{\beta}_1 = 2.088$	0.173	12.072
		$\widehat{\beta}_0 = -1.410$		
3	Tabaquismo	$\widehat{\beta}_1 = 0.069$	0.179	0.385
	Edad	$\widehat{\beta}_2 = 2.081$	0.174	11.956

con el mismo perfil, pero con la variable X_i que cambia en una unidad. Por ejemplo, en el primer modelo se tiene que $\widehat{\beta}_1 = 0.310$, de donde $exp[\widehat{\beta}_1] = 1.36$ es la estimación de la razón de *odds* correspondiente al hábito de fumar, el cual, nuevamente, se interpreta como el riesgo relativo de muerte ante al hábito de fumar.

Para el tercer modelo, $exp[\widehat{\beta}_2] = exp[2.081] = 8.01$ estima la razón de *odds* relacionada con la edad una vez que se controla el hábito de fumar (es decir, en tabaquismo = 0 o 1).

Ahora considere que se quieren emplear los datos de la tabla 5.21 para estimar $P(Y = 1)$ en sujetos de quienes se conoce la edad y el hábito de fumar. Para tal efecto se debe calcular

$$\frac{1}{1 + \exp[-\widehat{\beta}_0^* - 0.069X_1 - 2.081X_2]}$$

donde X_1 es el hábito de fumar ($X_1 = 0$ o 1) y X_2 la edad ($X_2 = 0$ o 1). El cálculo de $\widehat{\beta}_0^*$ se hace conforme a (5.19); como $\widehat{\beta}_0 = -1.41$ y $f = 0.2$

(pues mueren 400 de cada 2.000), de (5.20) se obtiene:

$$\begin{aligned}\widehat{\beta}_0^* &= \widehat{\beta}_0 - ln\left(\frac{n_1}{n_2}\frac{1-f}{f}\right)\\ &= -1.41 - ln\left(\frac{400}{400}\cdot\frac{0.8}{0.2}\right)\\ &= -1.41 - ln(4)\\ &= -2.79\end{aligned}$$

Se observa que este valor coincide con el obtenido para el modelo prospectivo (intercepto del modelo 3 de la tabla 5.14); con esto se puede afirmar que la transformación (5.20) corrige el efecto de no representatividad de la muestra.

5.9 Procesamiento de datos con R

A continuación se introducen los datos de la tabla 5.1 y se calculan los valores que aparecen en la sección 5.2.

Modelo de regresión logística

```
# introducción de la columna edad
edad<-c(34,21,54,67,32,56,76,44,34,21,48,39,22,45,
        65,67,22,32,21,76,45,23,44,65,66,74,34,43,
        47,37,26,54,53,55,23,34,43,45,31,55)
# Introducción de la columna infección
infeccion<-c("No","No","Si","No","Si","Si","Si","No",
             "No","No","No","No","No","No","Si","Si",
             "No","No","Si","Si","No","No","No","Si",
             "Si","Si","No","No","No","No","No","No",
             "No","Si","No","No","No","No","No","No")
#Se convierte infección a factor con 'Si' como nivel
#de referencia
infeccion<-relevel(factor(infeccion),ref="Si")
# introducción de la columna atención
atencion<-rep(c(0,1),c(20,20))
# por último se organizan los datos en un data.frame
```

```
t5.1<-data.frame(edad=edad,infeccion=infeccion,
                 atencion=atencion)
```

Los comandos que siguen proporcionan la tabla 5.2 y la prueba de independencia de los dos factores que equivalen a probar la hipótesis que las tasas de infección no difieren significativamente en los dos tipos de tratamientos.

```
t5.2<-table(t5.1$atencion,t5.1$infeccion)
# tabla 5.2
addmargins(t5.2)
# prueba ji--cuadrado
chisq.test(t5.2,correct = FALSE)
# equivalente a
summary(t5.2)
```

La prueba paramétrica de la hipótesis que la edad media del grupo infectado es igual a la edad media del grupo no infectado se realiza mediante el comando

```
with(t5.1,
t.test(edad~infeccion,var.equal=TRUE))
```

mientras que la prueba NO paramétrica de la misma hipótesis se realiza con

```
with(t5.1,
wilcox.test(edad~infeccion,correct=FALSE,exact=FALSE))
```

El modelo de regresión logística 5.5 se ajusta de la siguiente forma:

```
# ajuste del modelo
mod1<-glm(infeccion~atencion+edad,
          family=binomial,data=t5.1)
# para mostrar los resultados del ajuste.
summary(mod1)
```

A continuación se obtiene la probabilidad de que un paciente de 54 años con atención tradicional resulte infectado:

```
# nuevos datos
nd<-data.frame(edad=54, atencion=0)
predict(mod1,newdata=nd,type ="response")
```

La obtención del riesgo relativo para un paciente de 50 años atendido con la metodología propuesta frente a uno de 55 años atendido por la metodología tradicional se obtiene así:

```
# datos del primer individuo
nd1<-data.frame(edad=50, atencion=1)
# datos del segundo individuo
nd2<-data.frame(edad=55, atencion=0)
# probabilidad de infección del primer individuo
px1<-predict(mod1,newdata=nd1,type ="response")
# odds del primer individuo
OX1<-px1/(1-px1)
# probabilidad de infección del primer individuo
px2<-predict(mod1,newdata=nd2,type ="response")
# odds del primer individuo
OX2<-px2/(1-px2)
# riesgo relativo
OX1/OX2
```

Cálculos para la sección 5.4

En este apartado se ajusta el modelo y se realizan los cálculos de la sección 5.4. Primero se introducen los datos de la tabla 5.3 y luego se ajusta el modelo discutido en dicha sección.

```
# enfermedad coronaria: 1 si desarrolló 0 si no
e.cor<-c(0,1,0,1,0,1,1,1,1,0,0,1,1,1,1,0,0,1,1,0,1,1,0,0,
         0,1,0,0,0,0,1,1,0,1,0,0,0,0,1,1,1,1,0,0,0,0,0,0,
         0,0,0,1,0,1,1,0,0,1,1,1)
e.cor<-factor(e.cor)
# edad en años
edad<-c(48,55,54,51,51,59,60,50,58,41,51,59,50,50,62,
        60,55,48,50,48,56,61,52,42,44,47,46,61,43,41,
        62,50,48,56,61,52,42,44,47,56,51,52,42,65,49,
```

```
        41,42,46,54,51,51,59,60,50,58,41,52,43,58,52)
# hábito de fumar: 1 si fuma 0 si no fuma
fuma<-c(1,1,0,1,0,0,1,0,1,1,0,0,1,1,1,0,0,1,1,0,1,1,
        0,1,1,1,1,0,0,1,1,0,0,1,0,1,1,1,0,1,1,1,1,0,
        1,1,1,0,0,1,0,0,0,1,1,1,0,0,1,1)
# tensión arterial
tas<-c(125,160,140,135,135,150,160,140,160,135,135,
       150,160,150,150,140,140,150,160,130,155,155,
       135,135,140,165,130,150,125,150,110,125,150,
       150,155,125,135,140,145,150,155,125,135,145,
       155,145,135,140,135,135,135,150,145,160,160,
       135,155,160,155,160)
# se organizan los datos en un data.frame
t5.3<-data.frame(e.cor=e.cor,edad=edad,fuma=fuma,
                 tas=tas)
# una 'mirada' a los datos
head(t5.3)
tail(t5.3)
summary(t5.3)
```

A continuación se ajusta el modelo de regresión logística y se imprime el resumen del ajuste.

```
# regresión logística
mod2<-glm(e.cor~edad + fuma + tas,family=binomial,data=t5.3)
options(digits=4)
summary(mod2)
```

La estimación de la probabilidad de que una persona de 55 años, fumador con tensión arterial sistólica 150, padezca la enfermedad coronaria se obtiene con el siguiente código:

```
# nuevos datos
nd1<-data.frame(edad=55, fuma=1, tas=155)
# probabilidad
px1<-predict(mod2,newdata=nd1,type ="response")
# odds
OX1<-px1/(1-px1)
```

La misma probabilidad para una persona con el mismo perfil, pero que no fuma, se obtiene mediante el código siguiente. Además, se obtiene el riesgo relativo de fumar manteniendo fijas las demás variables:

```
# nuevos datos
nd2<-data.frame(edad=55, fuma=0, tas=155)
# probabilidad
px2<-predict(mod2,newdata=nd2,type ="response")
# odds
OX2<-px2/(1-px2)
# riesgo relativo de una persona que fuma frente
# a una que no
OX1/OX2
# otra forma de obtener el valor anterior es mediante
exp(coef(mod2)[3])
```

Por último, se obtienen las estadísticas básicas de las variables explicativas para cada nivel de la variable respuesta (enfermedad coronaria) y se elabora la tabla 5.4.

```
# variable edad, media
tapply(t5.3$edad , t5.3$e.cor, mean)
# desviación estándar
tapply(t5.3$edad , t5.3$e.cor, sd)
# media y desviación estándar global
mean(t5.3$edad)
sd(t5.3$edad)
#variable: fuma
tapply(t5.3$fuma   , t5.3$e.cor, mean )
tapply(t5.3$fuma  , t5.3$e.cor, sd)
mean(t5.3$fuma)
sd(t5.3$fuma)
# variable tas
by(t5.3$tas  , t5.3$e.cor, mean )
by(t5.3$tas  , t5.3$e.cor, sd)
tapply(t5.3$tas  , t5.3$e.cor, mean )
tapply(t5.3$tas  , t5.3$e.cor, sd)
mean(t5.3$tas)
sd(t5.3$tas)
```

Construcción de la tabla 5.4, para lo cual se requiere calcular los coeficientes estandarizados mediante la ecuación 5.10b. Se requiere la librería MASS para obtener los intervalos de confianza.

```
# coeficientes estandarizados
de<-apply(t5.3[,-1],2,sd)
betas<-coef(mod2)[-1]
sqrt(3)/pi*betas*de
# razones de odds
round( exp(coef(mod2)), 4)
library(MASS) # si ya se ha cargado no es necesario
# intervalos de confianza para las razones de  odds
exp(confint(mod2))
# construcción de la tabla
cbind( summary(mod2)$coef[,1],
round(summary(mod2)$coef[,2],4),
c(0,sqrt(3)/pi*betas*de) ,
round( exp(coef(mod2)), 4),
exp(confint(mod2)) )
```

5.9.1 Cálculos para la sección 5.6

En las líneas de código de R que siguen a continuación se calcula la estadística G^2 tratada en la sección 5.6.1, para lo cual se retoma el ejemplo de las enfermedades coronarias, cuyos datos se consignan en la tabla 5.3, que se introducen en el código presentado en el apartado anterior; por tanto, en esta oportunidad, no se hace la lectura de los datos.

Las órdenes siguientes buscan verificar si se justifica tener en el modelo las variables EDAD, FUMAR y TAS, en conjunto.

```
# ajuste del modelo completo
mod2<-glm(e.cor~edad + fuma + tas,family=binomial,
          data=t5.3)
# ajuste del modelo que solo incluye el intercepto.
mod21<-glm(e.cor~1,family=binomial,data=t5.3)
# comparación de los dos modelos (G^2)
anova(mod21,mod2,test="Chisq")
```

A continuación se decide entre el modelo que solo incluye la variable TAS y el modelo que tiene las tres variables.

```
# ajuste del modelo que solo incluye el TAS
mod21<-glm(e.cor~1,family=binomial,data=t5.3)
# comparación de los dos modelos (G^2)
anova(mod21,mod2,test="Chisq")
```

por último se imprime la tabla 5.6.

```
summary(mod2)
```

Cálculos para la sección 5.7

En este apartado se muestra el código de R para llevar a cabo el ajuste de modelos de regresión logística con respuesta politómica. Se hace para los dos casos tratados en la sección 5.7: regresión logística con respuesta politómica nominal (sección 5.7.1) y regresión logística con respuesta politómica ordinal (sección 5.7.2).

Regresión logística con respuesta politómica nominal

Para llevar a cabo el ajuste de este tipo de modelos se usan las librerías `epicalc`[14] y `nnet`[15]. Para mayores detalles sobre la librería `epicalc`, el lector puede consultar a Virasakdi (2004).

```
# se cargan las librerías
library(epicalc)
library(nnet)
# se introducen los datos
cont<-matrix(c(16,5,6, 6,7,19, 5,2,7, 1,0,10),ncol=3,
             byrow=TRUE)
colnames(cont)<-c("Bueno","Regular","Ninguno")
sexo<-rep(c("Femenino","Masculino"),c(2,2))
sexo<-factor(sexo)
```

[14]Epidemiological calculator.
[15]Neural Networks and Multinomial Log-Linear Models.

```
# se define Masculino como nivel de referencia
sexo<-relevel(sexo,ref="Masculino")
trat<-rep( c("Actividad","Placebo"),2)
trat<-factor(trat)
# se define Placebo como nivel de referencia
trat<-relevel(trat,ref="Placebo")
poli<-multinom(cont ~ sexo+trat); poli
summary(poli)
# para mostrar los RR e intervalos de confianza
mlogit.display(poli)
# probabilidades estimadas
data.frame(sexo,trat, fitted.values(poli))
```

Regresión logística con respuesta politómica ordinal

En el ajuste de los modelos de regresión logística politómica con respuesta ordinal se usa la función `polr()` de la librería MASS.

```
# introducción de los datos
cont<-c(16,5,6, 6,7,19, 5,2,7, 1,0,10)
sexo<-factor(rep(c("Femenino","Masculino"),c(6,6)))
# se define Masculino como nivel de
# referencia para sexo
sexo<-relevel(sexo,ref="Masculino")
trat<-factor(rep(rep(c("Actividad","Placebo"),
            c(3,3) ),2))
# se define Placebo como nivel de referencia
# para trat
trat<-relevel(trat,ref="Placebo")
mejoria<-factor( rep(c("Bueno","Regular",
                "Ninguno"),4) )
#### (importante) ####
# se define mejoria como un factor ordenado
# "Ninguno"<"Regular"<"Bueno"
mejoria<-ordered(mejoria,levels=c("Ninguno",
                "Regular","Bueno") )
# se organizan los datos en un data.frame
dft5.9<-data.frame(cont=cont,sexo=sexo,trat=trat,
                  mejoria=mejoria)
```

```
# se ajusta el modelo
poli2<-polr(mejoria~trat+sexo,weights=cont,
           data=dft5.9)
# se imprimen los resultados del ajuste
summary(poli2)
# probabilidades estimadas.
unique(data.frame(sexo,trat,fitted.values(poli2)))
```

5.10 Ejercicios

1. A un grupo de 30 pacientes se les proporcionó un agente anestésico que fue mantenido a cierta concentración durante 15 minutos antes de hacer una incisión. Se observó si el paciente se movió o se quejó al hacer la incisión. El interés es estimar cómo varía la probabilidad de moverse o quejarse al incrementar la concentración del agente anestésico. Los datos se muestran en la tabla 5.22.

Tabla 5.22: Pacientes que se mueven o quejan al hacer una incisión 15 minutos después de aplicada la concentración del anestésico.

Concentración	0.8	1	1.2	1.4	1.6	2.5
Se movieron o se quejaron	6	4	2	2	0	0
No se movieron ni se quejaron	1	1	4	4	4	2
Total	7	5	6	6	4	2

   a) Ajuste el modelo $\text{logit}(\pi) = \beta_0 + \beta_1 x$, donde π es la probabilidad de que un paciente no se mueva o no se queje cuando ha sido tratado con una concentración x.

   b) Discuta la significación del coeficiente β_1 usando el estadístico de Wald.

   c) Obtenga un intervalo de confianza al 95% para β_1 e interprételo.

   d) Estime la probabilidad de que un paciente al que se le aplicó una concentración de 2 se queje o se mueva.

   e) ¿Qué concentración debe aplicarse para que solo el 5% de los pacientes se muevan o se quejen al hacer una incisión?

2. La tabla 5.23 muestra el número de veces que ocurrió inhibición (no hubo flujo de corriente a través de una membrana) para diferentes concentraciones de la proteína *peptide–C*. El resultado *Sí* indica que ocurrió inhibición. Use la regresión logística para modelar la probabilidad de inhibición como una función de la concentración de proteína.

Tabla 5.23: Datos de inhibición.

Concentración	0.1	0.5	1	10	20	30	50	70	80	100	150
No	7	1	10	9	2	9	13	1	1	4	3
Sí	0	0	3	4	0	6	7	0	0	1	7

3. Se alimentó a animales de laboratorio con Aflatoxin B1 en diferentes dosis y se registró el número de casos con cáncer de hígado. Los datos se pueden obtener del marco de datos `aflatoxin` de la librería `faraway`.

   a) Construya un modelo para predecir la ocurrencia de cáncer de hígado. Calcule el nivel DL50 (dosis letal que mata al 50% de los expuestos).

   b) Discuta la utilidad del modelo para extrapolar a bajas dosis.

4. El marco de datos `esoph` de `R` contiene el resultado de un estudio caso–control sobre cáncer de esófago en Ileet–Vilaine, Francia. Los datos se pueden obtener con el comando `data(esoph)` y una descripción de las variables mediante `help(esoph)`.

   a) Ajuste un modelo logístico que incluya interacción entre los tres predictores. Use eliminación *backward* para simplificar el modelo tanto como sea razonable.

   b) ¿Cómo se ajusta el modelo final a los datos? Explique.

   c) ¿Cuál es la probabilidad de que una persona de 25 años que no fuma o toma desarrolle un cáncer de esófago?

5. El Instituto Nacional de Diabetes, Enfermedades Digestivas y del Riñón realizó un estudio que involucró a 768 mujeres indígenas adultas de la comunidad Pima que habita cerca de Phoenix. El propósito del estudio era investigar factores relacionados con la

diabetes. Los datos se pueden encontrar en el marco de datos `pima` de la librería `faraway`.

a) Realice un resumen gráfico y numérico simple de los datos.

b) Ajuste un modelo con los resultados de la prueba de diabetes como respuesta y todas las otras variables como predictoras. ¿Es bueno el ajuste del modelo? Justifique.

c) ¿Cuál es la diferencia en los *odds* de prueba positiva para diabetes en una mujer con BMI en el primer cuantil comparado con una mujer en el tercer cuantil, asumiendo que los demás factores se mantienen constantes?

d) Elabore un diagnóstico del modelo, reportando cualquier violación potencial de los supuestos y sugiriendo formas de mejorar el ajuste del modelo.

Capítulo 6

Análisis discriminante

6.1 Introducción

Dos son los objetivos principales abordados por el *análisis discriminante*: la *separación o discriminación de grupos*, y la *predicción* o *asignación* de un objeto en uno de entre varios grupos previamente definidos, con base en los valores de las variables que lo identifican. El primer objetivo es descriptivo; trata de encontrar las diferencias entre dos o más grupos a través de una *función discriminante*. En este capítulo se trata sobre el *análisis de clasificación*, el cual se orienta a "ubicar" un objeto o unidad muestral en uno de varios grupos de acuerdo con una *regla de clasificación* (o *regla de localización*). Sin embargo, frecuentemente la mejor función para separar grupos provee también la mejor regla de localización de observaciones futuras, de forma que estos dos términos generalmente se emplean indistintamente.

Por ejemplo:

- Conocer si un individuo va a desarrollar o no una enfermedad en función de una serie de variables analíticas conocidas.
- Clasificar a las familias de acuerdo con el hecho de que vayan a tener un enfermo terminal en su domicilio, a través de una serie de variables conocidas.
- En taxonomía, un biólogo quiere clasificar una "nueva" planta en una de varias especies conocidas.

- Un arqueólogo debe ubicar a un antepasado en uno de cuatro periodos históricos.
- En medicina forense, se quiere determinar el sexo que tenía un individuo de acuerdo con algunas medidas registradas en algunos de sus huesos.

Estos son algunos casos típicos de los que se ocupa el análisis discriminante pues, de acuerdo con un conjunto de variables, se quiere obtener una función con la cual se pueda decidir sobre la asignación de un caso (sujeto o individuo) a una de varias poblaciones mutuamente excluyentes.

En el análisis discriminante se obtiene una función que separa entre varios grupos definidos a priori, esta función es una suma ponderada de las variables de identificación, la cual minimiza los errores de clasificación. El problema de la discriminación es, entonces, comprobar si tales variables permiten diferenciar las clases definidas previamente y precisar cómo se puede hacer.

Cabe resaltar que el problema es *identificar* la clase (grupo o población) a la que se debe asignar un individuo, de quien se sabe que pertenece a una de las clases definidas de antemano, y para el cual solo se conocen los valores de las variables "explicativas". Se sigue entonces una tarea de discriminación descriptiva en primer lugar, con la que se asignan individuos a las clases, más no se agrupan, puesto que no se trata de construir grupos sino de asignar individuos a estos. La última característica diferencia la técnica de discriminación con la de clasificación (el análisis de conglomerados); otro tema es el empleo de estas técnicas para complementar o confrontar los resultados de una clasificación.

6.2 Reglas de discriminación para dos grupos

La mayor parte de la literatura sobre análisis discriminante trata el problema para dos poblaciones. Sobre la base de una información, contenida en un vector X de variables medidas sobre una unidad de observación (sujetos), que en adelante se indicará como la observación X, se quiere clasificar esta unidad en una de las dos poblaciones G_1 o G_2.

6.2.1 Vía máxima verosimilitud

Aunque esta situación, en la práctica, es muy poco frecuente, suponga que se conocen las distribuciones de las dos poblaciones. Sean $f_1(X)$ y $f_2(X)$ las *fdp* de cada una de las poblaciones, con X vector de observaciones de tamaño $(p \times 1)$ (un caso, sujeto o individuo). La regla de discriminación *máximo verosímil* para localizar el caso caracterizado por X en alguna de las dos poblaciones consiste en ubicarlo en la población para la cual X maximiza la verosimilitud o probabilidad.

En símbolos, si G_1 y G_2 son las dos poblaciones, entonces se localiza a X en G_i si

$$L_i(X) = \max_j\{L_j\}, \quad \text{con } i, j = 1, 2 \tag{6.1}$$

Aunque parece una perogrullada, la regla sencillamente sugiere que el caso se asigne a la población para la cual la probabilidad de proceder de ella sea máxima. La regla dada en (6.1) es extensible a cualquier número de poblaciones. En caso de empates, X se asigna a cualquiera de las poblaciones.

Poblaciones con matrices de covarianzas iguales

Para la construcción de la regla de asignación se asume de entrada que los dos grupos (clases o poblaciones) tienen distribución normal multivariada con la misma matriz de covarianzas.[1]

Se extrae la muestra $X_{1(i)}, \ldots, X_{n_i(i)}$ de $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$[2] para $i = 1, 2$. Esto es equivalente a tener dos matrices de datos, una por cada muestra en la respectiva población, así:

Muestra 1				Muestra 2			
$X_{1_1 1}$	$X_{1_1 2}$	$\cdots$	$X_{1_1 p}$	$X_{1_2 1}$	$X_{1_2 2}$	$\cdots$	$X_{1_2 p}$
$X_{2_1 1}$	$X_{2_1 2}$	$\cdots$	$X_{2_1 p}$	$X_{2_2 1}$	$X_{2_2 2}$	$\cdots$	$X_{2_2 p}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X_{n_1 1}$	$X_{n_1 2}$	$\cdots$	$X_{n_1 p}$	$X_{n_2 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_2 p}$

[1] En el caso multivariado se tiene, además de la variabilidad de cada variable, la correlación entre estas.

[2] Esta notación hace referencia a la distribución normal multivariada.

Además de la información anterior se debe incluir una variable indicadora (como se insinúa con los subíndices en las matrices de datos) que señale la población a la que pertenece cada caso u observación.

Con base en esta información se pretende asignar la observación X a G_1 o a G_2. Los estimadores para el vector de medias $\boldsymbol{\mu}_i$ y la matriz de covarianzas $\boldsymbol{\Sigma}$ son, respectivamente:

$$\bar{X}_i = \sum_{j=1}^{n_i} X_{j(i)}/n_i,\ i = 1, 2,$$

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}\Big[\sum_{j=1}^{n_1}(X_{j(1)} - \bar{X}_1)(X_{j(1)} - \bar{X}_1)' + \sum_{j=1}^{n_2}(X_{j(2)} - \bar{X}_2)(X_{j(2)} - \bar{X}_2)'\Big].$$

Con estos datos muestrales los criterios de asignación son

$$\begin{array}{ll} \text{si} & \widehat{b}'X \geq \widehat{b}'\bar{X}_c, \quad X \text{ se asigna a } G_1 \text{ o,} \\ \text{si} & \widehat{b}'X < \widehat{b}'\bar{X}_c, \quad X \text{ se asigna a } G_2 \end{array} \tag{6.2}$$

con $\bar{X}_c = \frac{1}{2}(\bar{X}_1 + \bar{X}_2)$ y $\widehat{b} = \mathbf{S}^{-1}(\bar{X}_1 - \bar{X}_2)$.

Igual que en regresión, los puntos $(\bar{X}_1, \bar{Y}_1)$ y $(\bar{X}_2, \bar{Y}_2)$ satisfacen la ecuación $Y = b'X$, es decir, $\bar{Y}_1 = b'\bar{X}_1$ y $\bar{Y}_2 = b'\bar{X}_2$, de manera que $Y_c = b'X_c$. Las decisiones contempladas en (6.2) son equivalentes a:

$$\begin{array}{ll} \text{si} & \widehat{Y} \geq \bar{Y}_c = \frac{\bar{Y}_1 + \bar{Y}_2}{2}, \quad X \text{ se asigna a } G_1, \text{ o} \\ \text{si} & \widehat{Y} < \bar{Y}_c = \frac{\bar{Y}_1 + \bar{Y}_2}{2}, \quad X \text{ se asigna a } G_2 \end{array} \tag{6.3}$$

con $\bar{Y}_c = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)$ y $\widehat{Y} = \widehat{b}X = \mathbf{S}^{-1}(\bar{X}_1 - \bar{X}_2)X$.

La figura (6.1) ilustra la discriminación entre dos grupos que tienen distribución normal bivariada a través de la función discriminante lineal estimada $\widehat{Y} = \widehat{b}'X$. Por la forma y escala de las gráficas las matrices de covarianzas se han supuesto iguales. Cuando la función se aplica en un punto $X_i = (X_{i1}, X_{i2})'$, se obtiene la combinación lineal $Y_i = b_1X_{i1} + b_{i2}$; Y_i corresponde a la proyección del punto X_i sobre la línea óptima para la separación de los dos grupos. Como las dos variables X_1 y X_2 son normales (pues X es normal bivariada), la combinación lineal de estas es nuevamente normal.

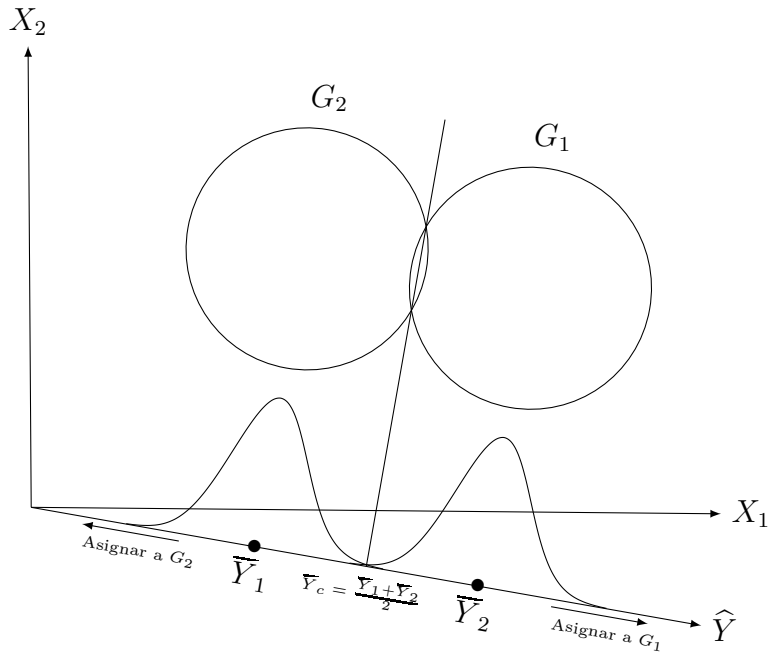

Figura 6.1: Discriminación lineal.

La magnitud del valor de la función de clasificación en el punto X respecto al punto $\overline{Y}_c$ define la asignación de la observación X a uno de los dos grupos.

A un grupo de 49 personas, de edad avanzada, que participaron en un estudio, se les clasificó mediante evaluación psiquiátrica en una de las dos categorías: *senil* o *no senil*. Una prueba de inteligencia adulta, independientemente administrada a cada uno, revela grandes diferencias entre los dos grupos en algunas partes de la prueba. Las medias de estas partes de la prueba se resumen en la tabla siguiente: La matriz de covarianzas muestral es

$$\mathbf{S} = \begin{pmatrix} 11.26 & 9.40 & 7.15 & 3.38 \\ 9.40 & 13.53 & 7.38 & 2.55 \\ 7.15 & 7.38 & 11.57 & 2.62 \\ 3.38 & 2.55 & 2.62 & 5.81 \end{pmatrix}$$

El estimador de la función de discriminación para una observación

Tabla 6.1: Evaluación psiquiátrica.

Variable	Subprueba	No senil ($n_1 = 37$)	Senil ($n_2 = 12$)
X_1	Información	12.57	8.75
X_2	Similaridades	9.57	5.33
X_3	Aritmética	11.49	8.50
X_4	Hábil. artist.	7.97	4.75

Fuente: Morrison (1990, 143).

$X' = (X_1, X_2, X_3, X_4)$ es

$$\begin{aligned}\widehat{Y} = \widehat{b}'X &= (\bar{X}_1 - \bar{X}_2)'\mathbf{S}^{-1}X \\ &= \begin{pmatrix}3.82 & 4.24 & 2.99 & 3.22\end{pmatrix}\begin{pmatrix}11.26 & 9.40 & 7.15 & 3.38\\ 9.40 & 13.53 & 7.38 & 2.55\\ 7.15 & 7.38 & 11.57 & 2.62\\ 3.38 & 2.55 & 2.62 & 5.81\end{pmatrix}^{-1}\begin{pmatrix}X_1\\X_2\\X_3\\X_4\end{pmatrix} \\ &= 0.030X_1 + 0.204X_2 + 0.010X_3 + 0.443X_4.\end{aligned}$$

Para ubicar un individuo en alguno de los dos grupos (senil o no senil) se utilizan los criterios expuestos en las desigualdades (6.2):

$$\bar{X}_c' = \frac{1}{2}(\bar{X}_1 + \bar{X}_2) = (10.66, 7.45, 9.99, 6.36),$$

como

$$\widehat{b}'X_c = (0.030, 0.204, 0.010, 0.443)(10.66, 7.45, 9.99, 6.36)' = 4.7512.$$

Se asigna un individuo al grupo *no senil* si la función de discriminación estimada $\widehat{Y}_i \geq 4.7512$, y a la categoría *senil* si $\widehat{Y}_i < 4.7512$ (figura (6.2)). Suponga que un individuo obtuvo los puntajes dados en el vector $X_0 = (10, 8, 7, 5)$, el valor de la función de discriminación en este caso es $\widehat{Y} = \widehat{b}'X_0 = 4.2115$: Como este valor es menor que 4.7512, el individuo es ubicado en el grupo senil.

Observación

Se nota alguna semejanza entre el modelo de regresión lineal y la función discriminante. Aunque en algunos cálculos son parecidos, estas técnicas tienen algunas diferencias estructurales:

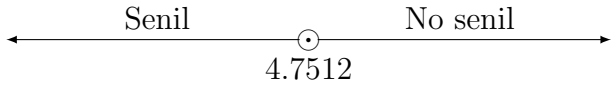

Figura 6.2: Discriminación en senil o no senil.

- En primer lugar, en el análisis de regresión se asume que la variable dependiente se distribuye normalmente y los regresores se consideran fijos. En análisis discriminante, la situación es al revés: las variables independientes se asumen distribuidas normalmente y la variable respuesta es fija, y toma los valores de cero o uno, según la ubicación del objeto en alguno de los dos grupos.

- En segundo término, el objetivo principal del análisis de regresión es predecir la respuesta media con base en el conocimiento de algunos valores fijos de un conjunto de variables explicativas; en cambio, el análisis discriminante pretende encontrar una combinación lineal de variables independientes que minimicen la probabilidad de clasificar incorrectamente objetos en sus respectivos grupos.

- Finalmente, el análisis de regresión propone un modelo formal, sobre el que se hacen ciertos supuestos con el fin de generar estimadores de los parámetros que tengan algunas propiedades deseables. El análisis discriminante busca un procedimiento para asignar o clasificar casos a grupos.

Poblaciones con matrices de covarianzas distintas

Si las dos poblaciones G_1 y G_2 tienen distribución normal p-variante con matrices de covarianzas distintas ($\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$), el logaritmo de la razón de

la verosimilitud para una observación particular X es el siguiente

$$\begin{aligned}
Q(X) &= \ln\left(\frac{L_1(X)}{L_2(X)}\right) \\
&= \ln\left(\frac{(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(X-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(X-\boldsymbol{\mu}_1)\}}{(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(X-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(X-\boldsymbol{\mu}_2)\}}\right) \\
&= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}(X-\boldsymbol{\mu}_1)'{\boldsymbol{\Sigma}_1}^{-1}(X-\boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2}(X-\boldsymbol{\mu}_2)'{\boldsymbol{\Sigma}_2}^{-1}(X-\boldsymbol{\mu}_2) \\
&= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}\big(\boldsymbol{\mu}_1'{\boldsymbol{\Sigma}_1}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'{\boldsymbol{\Sigma}_2}^{-1}\boldsymbol{\mu}_2\big) \\
&\quad + \big(\boldsymbol{\mu}_1'{\boldsymbol{\Sigma}_1}^{-1} - \boldsymbol{\mu}_2'{\boldsymbol{\Sigma}_2}^{-1}\big)X - \frac{1}{2}X'\big({\boldsymbol{\Sigma}_1}^{-1} - {\boldsymbol{\Sigma}_2}^{-1}\big)X. \qquad (6.4)
\end{aligned}$$

En la expresión $Q(X)$, el último término, $X'\big({\boldsymbol{\Sigma}_1}^{-1} - {\boldsymbol{\Sigma}_2}^{-1}\big)X$, corresponde a los cuadrados y productos cruzados de las componentes del vector X; $Q(X)$ se denomina *función de discriminación cuadrática*. Note que si $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, se tiene la función de discriminación lineal.

Se asigna la observación X a la población o grupo G_1 si $Q(X) \geq 0$; en caso contrario se asigna a la población G_2.

En términos muestrales, si se obtiene una muestra de la población G_1 y una de la población G_2, se calcula un valor muestral de $\widehat{Q}(X)$ al remplazar $\boldsymbol{\mu}_i$ por $\bar{X}_i$ y $\boldsymbol{\Sigma}_i$ por $\boldsymbol{S}_i$ (i-ésimo grupo). Esta es:

$$\begin{aligned}
\widehat{Q}(X) =& \frac{1}{2}\ln\left(\frac{|\boldsymbol{S}_2|}{|\boldsymbol{S}_1|}\right) - \frac{1}{2}\big(\bar{X}_1'{\boldsymbol{S}_1}^{-1}\bar{X}_1 - \bar{X}_2'{\boldsymbol{S}_2}^{-1}\bar{X}_2\big) \\
&+ \big(\bar{X}_1'{\boldsymbol{S}_1}^{-1} - \bar{X}_2'{\boldsymbol{S}_2}^{-1}\big)X - \frac{1}{2}X'\big({\boldsymbol{S}_1}^{-1} - {\boldsymbol{S}_2}^{-1}\big)X. \qquad (6.5)
\end{aligned}$$

Se observa que $\widehat{Q}(X)$ tiene forma cuadrática, la cual se expresa en forma general como

$$\widehat{Q}(X) = b + c'X - X'\boldsymbol{A}X.$$

La regla para clasificar una observación muestral X es similar al caso poblacional; es decir, se asigna la observación o individuo X al grupo G_1 si $\widehat{Q}(X) \geq 0$; y al grupo G_2 en caso contrario.

Cuando $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, la función de clasificación cuadrática $\widehat{Q}(X)$ es óptima de manera asintótica. Para muestras de tamaño grande y con amplias

diferencias entre $\mathbf{\Sigma}_1$ y $\mathbf{\Sigma}_2$, la función de discriminación cuadrática es la más recomendable.

6.3 Reglas de discriminación para varios grupos

Se considera el caso de muestras obtenidas de k grupos $G_1, G_2, \ldots, G_k$. Se desarrollan reglas de discriminación para el caso de varias poblaciones que tienen matrices de covarianzas igual o distinta, respectivamente.

6.3.1 Grupos con matrices de covarianzas iguales

Cuando se muestrea, varias poblaciones normales con matrices de covarianzas iguales, las funciones de discriminación óptima son lineales. Estas funciones de clasificación se obtienen aquí.

Si $p_1, p_2, \ldots, p_k$ son las probabilidades a priori de que una observación X proceda de la población $G_1, G_2, \ldots, G_k$, respectivamente, la regla de clasificación óptima, conociendo las funciones de densidad, es la siguiente:

Asignar X a la población G_i si $p_i f_i(X) \geq p_j f_j(X)$, para todo $j = 1, \ldots, k$; es decir, como en la ecuación (6.1), $p_i f_i(X) = \text{máx}_j\{p_j f_j(X)\}$. Maximizar $p_i f_i(X)$ es equivalente a maximizar $\ln(p_i f_i(X))$. Si $X \sim N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, se obtiene

$$\ln(p_i f_i(X)) = \ln(p_i) - \frac{1}{2} p \ln(2\pi) - \frac{1}{2} \ln|\mathbf{\Sigma}| - \frac{1}{2}(X - \boldsymbol{\mu}_i)' \mathbf{\Sigma}^{-1} (X - \boldsymbol{\mu}_i)$$

donde $\mathbf{\Sigma}$ es la varianza común a las k poblaciones. Note que cuando no hay información a priori sobre las p_i, se opta por asumir que estas son iguales (distribución no informativa), y estas cantidades p_i desaparecen de la regla de clasificación; la regla de discriminación es entonces la de *máxima verosimilitud*. Además, se debe advertir que p es el número de variables, mientras que los p_i son las probabilidades a priori. Al desarrollar los cálculos algebraicos sobre la expresión anterior, se obtiene

$$\ln(p_i) + \boldsymbol{\mu}_i' \mathbf{\Sigma}^{-1} X - \frac{1}{2} \boldsymbol{\mu}_i' \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_i$$

Para observaciones muestrales (n_i por grupo), se asigna la observación X al grupo para el cual se maximice

$$\mathcal{D}_i = \ln(p_i) + \bar{X}_i' \boldsymbol{S}_p{}^{-1} X - \frac{1}{2} \bar{X}_i' \boldsymbol{S}_p{}^{-1} \bar{X}_i, \tag{6.6}$$

donde

$$\boldsymbol{S}_p = \frac{\sum_{i=1}^{k} (n_i - 1) \boldsymbol{S}_i}{\sum_{i=1}^{k} (n_i - 1)}.$$

Si se asume igual probabilidad a priori (p_i), entonces la observación X se asigna al grupo G_i que produzca el mayor valor $\mathcal{D}_i$. Alternativamente, se puede definir $\mathcal{D}_{ij} = \mathcal{D}_i - \mathcal{D}_j$, tal que la regla de asignación, por ejemplo para $i = 1, 2, 3$, es:

- Asignar a G_1 si $\mathcal{D}_{12} > 0$ y $\mathcal{D}_{13} > 0$
- Asignar a G_2 si $\mathcal{D}_{12} < 0$ y $\mathcal{D}_{23} > 0$
- Asignar a G_3 si $\mathcal{D}_{13} < 0$ y $\mathcal{D}_{23} < 0$

De esta manera, el espacio de los individuos es dividido en tres regiones de discriminación, cuyas fronteras vienen dadas por las reglas de asignación $\mathcal{D}_{ij}$. En la figura 6.3 se muestran estas tres regiones de discriminación para el caso de dos variables, X_1 y X_2 ($p = 2$) y tres grupos ($k = 3$).

6.3.2 Grupos con matrices de covarianzas distintas

Si se emplea la función de discriminación lineal para grupos con matrices de covarianzas distintas, las observaciones tienden a ser clasificadas en los grupos que tienen varianzas altas. De todas formas, la regla de clasificación puede modificarse conservando de manera óptima la clasificación, en términos de los errores de clasificación.

Considerando k-poblaciones de p-variables cada una, con distribución $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, y cada una con probabilidades a priori $p_1, \ldots, p_k$, respectivamente, se tiene:

$$\ln[p_i f(X|G_i)] = \ln(p_i) - \frac{1}{2} p \ln(2\pi) - \frac{1}{2} |\boldsymbol{\Sigma}_i| - \frac{1}{2} (X - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (X - \boldsymbol{\mu}_i)$$

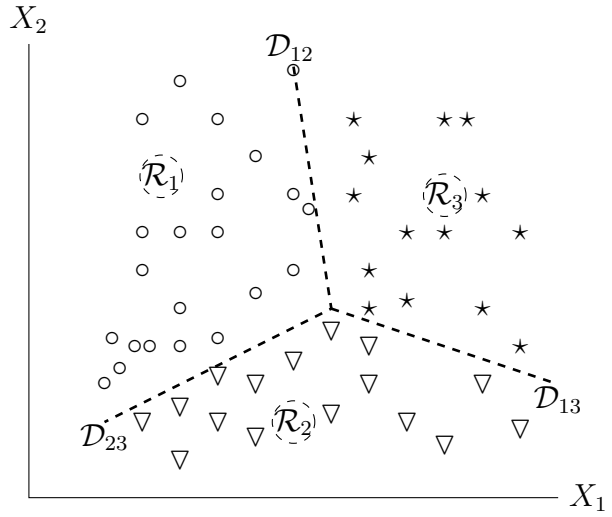

Figura 6.3: Regiones de discriminación para tres grupos.

Para una muestra se emplea el vector de medias muestral $\bar{X}_i$ y la matriz de covarianzas muestral $\boldsymbol{S}_i$, para cada uno de los k-grupos. Omitiendo el término constante $-(p/2)\ln(2\pi)$, se obtiene la función de discriminación cuadrática

$$\begin{aligned} Q_i(X) &= \ln(p_i) - \frac{1}{2}|\boldsymbol{S}_i| - \frac{1}{2}(X-\bar{X}_i)'\boldsymbol{S}_i^{-1}(X-\bar{X}_i) \\ &= \ln(p_i) - \frac{1}{2}|\boldsymbol{S}_i| - \frac{1}{2}\bar{X}_i'\boldsymbol{S}_i^{-1}\bar{X}_i + \frac{1}{2}\bar{X}_i'\boldsymbol{S}_i^{-1}X - \frac{1}{2}X'\boldsymbol{S}_i^{-1}X. \quad (6.7) \end{aligned}$$

La regla de clasificación es: asignar la observación X al grupo para el cual $Q_i(X)$ sea la más grande. Si las probabilidades a priori p_i son iguales o no se conocen, el término $\ln(p_i)$ puede descartarse de la función de discriminación. Note que para que exista $\boldsymbol{S}_i$, se debe satisfacer que $n_i > p$, con $i = 1, \ldots, k$.

Para poblaciones multinormales con matrices de covarianzas desiguales $\boldsymbol{\Sigma}_i$, la probabilidad a posteriori (bayesiana), empleando los estimadores de $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$, está dada por

$$P(G_i|X) = \frac{p_i|\boldsymbol{S}_i|^{-\frac{1}{2}}\exp\{-\frac{1}{2}D_i^2\}}{\sum_{i=1}^{k} p_i|\boldsymbol{S}_i|^{-\frac{1}{2}}\exp\{-\frac{1}{2}D_i^2\}}, \quad (6.8)$$

donde $D_i^2 = (X-\bar{X}_i)'\boldsymbol{S}_i^{-1}(X-\bar{X}_i)$. Aunque en la mayoría de las aplicaciones no se tienen los valores de p_i, algunos paquetes estadísticos

los estiman como una proporción de los tamaños de muestra n_i; este procedimiento no es muy recomendado, a menos que las proporciones muestrales representen las proporciones poblacionales.

Para tamaños muestrales grandes, la función de discriminación cuadrática clasifica mejor que las lineales. Para muestras de tamaño pequeño, los resultados desde la discriminación cuadrática son menos estables en muestreos secuenciales o repetitivos que los resultados de la discriminación lineal, pues se deben estimar más parámetros en $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_k$ que en $\boldsymbol{S}_p$ y porque cada $\boldsymbol{S}_i$ tiene asociado algunos pocos grados de libertad de $\boldsymbol{S}_p$.

6.4 Tasas de error de clasificación

Una vez obtenida una regla de clasificación, la inquietud natural es acerca de *qué tan buena* es la clasificación generada a través de esta regla. Es decir, se quiere saber la *tasa de clasificación correcta*, referida como la probabilidad de clasificar una observación en el grupo al que verdaderamente pertenece. De manera complementaria, se tienen las *tasas de error* por clasificación incorrecta. El interés está en la probabilidad que la regla de discriminación disponible clasifique incorrectamente una futura observación; de otra forma, se quiere evaluar la capacidad de la regla para predecir el grupo a que pertenece una observación. La siguiente tabla ilustra la calidad de las posibles decisiones que se puedan tomar con relación a la clasificación de objetos en uno de dos grupos.

Decisión estadística

Grupo	G_1	G_2
G_1 (n_1)	Decisión correcta (n_{11})	Error (n_{12})
G_2 (n_2)	Error (n_{21})	Decisión correcta (n_{22})

6.4.1 Estimación de las tasas de error

Un estimador simple de la tasa de error se obtiene al tratar de clasificar los objetos del mismo conjunto empleado para la construcción de la regla de clasificación. Este método se conoce como *resustitución*. A cada observación X_i se le aplica la función de clasificación y se asigna a uno de los grupos. Se cuentan entonces el número de clasificaciones correctas y el número de clasificaciones incorrectas conseguidas con la regla. La proporción de clasificaciones incorrectas se denomina *tasa de error aparente*. Los resultados se disponen en un cuadro de decisión como el anterior.

Entre las n_1 observaciones de G_1, n_{11} son clasificadas correctamente en G_1 y n_{12} son clasificadas incorrectamente en G_2, con $n_1 = n_{11} + n_{12}$. Análogamente, de las n_2 observaciones de G_2, n_{21} son asignadas incorrectamente a G_1 y n_{22} son correctamente asignadas a G_2, con $n_2 = n_{21} + n_{22}$. De esta forma, la tasa de error aparente es:

$$\begin{aligned} \textit{Tasa de error aparente} &= \frac{n_{12} + n_{21}}{n_1 + n_2} \\ &= \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}. \end{aligned} \tag{6.9}$$

El método de resustitución puede extenderse al caso de varios grupos.

La tasa de error aparente es fácil de calcular, aunque la mayoría de los paquetes estadísticos la suministran. Esta tasa es un estimador de la probabilidad que la función de clasificación, encontrada desde los datos, clasifique incorrectamente una observación. Tal probabilidad se denomina *tasa actual de error*. Si p_1 y p_2 son las probabilidades a priori para los grupos G_1 y G_2, respectivamente, la tasa actual de error es:

$$\begin{aligned} \textit{Tasa actual de error} = p_1 P(\textit{Asignar a } G_1 | G_2) \\ + p_2 P(\textit{Asignar a } G_2 | G_1), \end{aligned} \tag{6.10}$$

con $P(\textit{Asignar a } G_1|G_2)$ significa la probabilidad de clasificar X en el grupo G_1 cuando realmente procede del grupo G_2, una definición análoga se tiene para $P(\textit{Asignar a } G_2|G_1)$.

La definición de tasa actual de error se estima para procedimientos de clasificación basados en una muestra. Aunque se puede estar interesado en calcular la tasa de error promedio basados sobre todas las posibles

muestras; es decir,

$$\begin{aligned}\textit{Tasa actual esperada de error} = p_1\,\mathrm{E}[P(\textit{Asignar a } G_1|G_2)] \\ + p_2\,\mathrm{E}[P(\textit{Asignar a } G_2|G_1)]. \quad (6.11)\end{aligned}$$

En el cálculo de (6.10) o de (6.11) se necesita conocer los parámetros poblacionales y asumir una distribución particular de los datos. Pero en la mayoría de los casos los parámetros poblacionales son desconocidos; por tanto, se requieren algunos estimadores de las tasas de error.

6.5 Otras técnicas de discriminación

6.5.1 Modelo de discriminación logística para dos grupos

Cuando las variables son discretas o son una mezcla de discretas y continuas, la discriminación a través del modelo logístico puede resultar adecuada.

Para distribuciones multinormales con $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, el logaritmo de la razón de densidades es

$$\begin{aligned}\ln \frac{f(X|G_1)}{f(X|G_2)} &= -\frac{1}{2}\underbrace{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}_{\boldsymbol{\alpha}} + \underbrace{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}}_{\boldsymbol{\beta}'} X \\ &= \boldsymbol{\alpha} + \boldsymbol{\beta}' X, \quad (6.12)\end{aligned}$$

la cual es una función lineal del vector observado X. Además de la normal multivariada, otras distribuciones multivariadas satisfacen (6.12), algunas de las cuales involucran vectores aleatorios discretos o mezcla de variables discretas y continuas. El modelo mostrado en la ecuación (6.12) se conoce como *modelo logístico*. La regla para asignar una observación X es: Asignar al grupo G_1 si

$$\boldsymbol{\alpha} + \boldsymbol{\beta}' X > \ln \frac{p_1}{p_2}, \quad (6.13)$$

y a G_2 en otro caso. Cuando las probabilidades a priori, p_1 y p_2, se pueden asumir iguales, el miembro izquierdo de la desigualdad (6.13) se compara contra el número cero. La *clasificación logística* se refiere también como la *discriminación logística*.

La probabilidad a posteriori, en términos del modelo logístico, que señala la probabilidad de pertenencia de una observación X a un grupo, por ejemplo G_1, de acuerdo con el teorema de Bayes es

$$\begin{aligned} P(G_1|X) &= \frac{p_1 f(X|G_1)}{p_1 f(X|G_1) + p_2 f(X|G_2)} \\ &= \frac{e^{\ln(p_1/p_2)+\alpha+\beta' X}}{1 + e^{\ln(p_1/p_2)+\alpha+\beta' X}} \\ &= \frac{e^{\alpha_0+\beta' X}}{1 + e^{\alpha_0+\beta' X}}, \end{aligned} \tag{6.14}$$

donde $\alpha_0 = \ln(p_1/p_2) + \alpha$. De la expresión anterior se obtiene

$$P(G_2|X) = 1 - P(G_1|X) = \frac{1}{1 + e^{\alpha_0+\beta' X}}. \tag{6.15}$$

La estimación de $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ se hace a través del método de mínimos cuadrados ponderados o mediante máxima verosimilitud para regresión logística. La estimación lleva a resolver sistemas de ecuaciones no lineales, cuya solución aproximada puede encontrarse con métodos numéricos como la técnica de Newton-Raphson, el método de cuasi-Newton; procedimientos incorporados en paquetes estadísticos como el SAS o `R`.

La figura (6.4) representa la función logística. Aquí se asigna la observación X al grupo G_1, si $P(G_1|X) \geq P(G_2|X)$, o al grupo G_2, en caso contrario. En general, para dos grupos, de acuerdo con la propiedad expresada en (6.15), se asigna la observación X al grupo G_i si $P(G_i|X) \geq 0.5$, $i = 1, 2$.

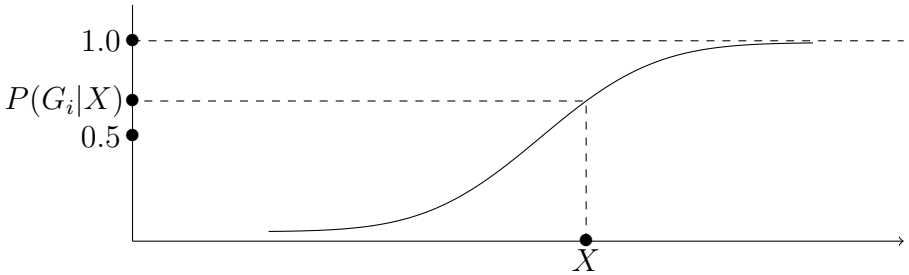

Figura 6.4: Función logística.

Para datos multinormales con $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, la clasificación lineal es superior a la logística; sin embargo, para datos binarios, estos supuestos usual-

mente no se tienen y la clasificación a través de un modelo logístico resulta mejor.

Los modelos de discriminación logística lineal pueden emplearse en situaciones donde:

1. Las funciones de densidad sean multinormales con matrices de covarianzas iguales.
2. Las mediciones sean variables independientes tipo Bernoulli.
3. Las variables tipo Bernoulli sigan un modelo log-lineal con efectos de segundo orden o más iguales.
4. Situaciones 1 a 3 mezcladas.

La clasificación logística se aplica, con buenos resultados, en investigación médica[3]. El objetivo es predecir una trombosis posoperatoria de venas profundas, una condición es que se debe tratar a estos pacientes con anticoagulantes antes de la cirugía. Sin embargo, estos tratamientos producen problemas hemorrágicos en algunos pacientes, de donde resulta importante la identificación de los pacientes con más alto riesgo de trombosis.

De 124 pacientes en estudio, ninguno mostró evidencia preoperatoria de trombosis en venas profundas. Después de la intervención, 20 pacientes desarrollaron la condición (grupo G_1) y los 104 restantes no (grupo G_2). En el modelo logístico resultante se consideran, finalmente, cuatro variables continuas (X_1, X_2, X_3, X_4) y una variable discreta X_5. El modelo (6.12) estimado es

$$\begin{aligned}\omega &= \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\beta}}' X \\ &= -11.3 + 0.009X_1 + 0.22X_2 + 0.043X_3 + 2.19X_4 + 0.085X_5.\end{aligned}$$

El valor de ω se calculó para cada uno de los 124 pacientes, remplazando por los respectivos valores de X_1 a X_5. Si se aplica la regla de clasificación (6.12), con $p_1 = p_2$, los pacientes con $\omega > 0$ se asignan al grupo de trombosis de venas profundas (G_1). Con este procedimiento, 11 de los 124 pacientes se clasificaron incorrectamente, es decir, con una tasa de error aparente de 9% (11/129). Sin embargo, usando el criterio de $\omega > 0$,

[3]Rencher (1998, 255-256).

se clasificaría incorrectamente a pacientes con alto riesgo (pues $p_1 > p_2$, para estos casos). Por tanto, se recomienda suministrar anticoagulante, antes de la cirugía, a los pacientes con $\omega > -2.5$.

En el caso de grupos con distribuciones multinormales donde $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, la función logística no es lineal en los X. El logaritmo de la razón de densidades es

$$\begin{aligned} \ln \frac{f(X|G_1)}{f(X|G_2)} &= c_0 + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})X + \frac{1}{2}X'(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})X \\ &= c_0 + \boldsymbol{\delta}'X + X'\boldsymbol{\Delta}X, \end{aligned} \tag{6.16}$$

con $c_0 = \frac{1}{2}\ln(|\boldsymbol{\Sigma}_2^{-1}|/|\boldsymbol{\Sigma}_1^{-1}|) - \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$, $\boldsymbol{\delta} = (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})$, y $\boldsymbol{\Delta} = (\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})$.

Aunque la función dada en (6.16) no es lineal en los X, es lineal en los parámetros. Esta se conoce como la función *logística cuadrática*. Los parámetros se estiman mediante los mismos métodos iterativos citados anteriormente.

6.5.2 Modelo de discriminación Probit

En algunos casos los grupos son definidos a través de un criterio *cuantitativo* en lugar de cualitativo. Por ejemplo, se puede particionar un grupo de estudiantes en dos grupos, con base en su promedio en el rendimiento académico, tal que en un grupo se ubican los de rendimiento "alto" y en el otro los de rendimiento "bajo". Sobre la base de un vector X de puntajes y medidas obtenidas para esta clase de estudiantes, se quiere predecir su pertenencia a uno de estos grupos.

A continuación se presentan los rasgos generales de la metodología. Sea Z una variable aleatoria continua. Si t es un valor "umbral" o "límite", entonces un individuo es asignado al grupo G_1; si $Z > t$ (por ejemplo, alto rendimiento) y si $Z \leq t$, se asigna al grupo G_2.

Para empezar se asume que el vector $(Z, X)'$ se distribuye $N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_Z \\ \boldsymbol{\mu}_X \end{pmatrix} \text{ y } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_Z^2 & \sigma_{ZX} \\ \sigma_{XZ} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}$$

Por propiedades de la distribución condicional en distribuciones multi-

normales, se tiene[4]:

$$\begin{aligned}\mathrm{E}(Z|X) &= \mu_{Z|X} = \mu_Z + \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X),\\ \mathrm{var}(Z|X) &= \sigma_{Z|X} = \sigma_Z^2 - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}\end{aligned}$$

Por tanto,

$$\begin{aligned}P(G_1|X) &= P(Z > t|X)\\ &= P\left(\frac{Z - \mu_{Z|X}}{\sigma_{Z|X}} > \frac{t - \mu_{Z|X}}{\sigma_{Z|X}}\right)\\ &= 1 - \Phi\left(\frac{t - \mu_{Z|X}}{\sigma_{Z|X}}\right)\\ &= \Phi\left(\frac{-t + \mu_{Z|X}}{\sigma_{Z|X}}\right),\end{aligned}$$

donde $\Phi(\cdot)$ es la función de distribución normal estándar. De esta forma, remplazando $\mu_{Z|X}$ y $\sigma_{Z|X}$, la probabilidad que la observación X sea del grupo G_1 es:

$$\begin{aligned}P(G_1|X) &= \Phi\left[\frac{-t + \mu_Z + \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X)}{\sigma_Z^2 - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}}\right]\\ &= \Phi(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X),\end{aligned} \tag{6.17}$$

donde

$$\boldsymbol{\gamma}_0 = -(t - \mu_Z + \sigma'_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X)/\sqrt{\sigma_Z^2 - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}},\ \text{y}$$

$$\boldsymbol{\gamma}_1 = \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}/\sqrt{\sigma_Z - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}}.$$

La regla de clasificación asigna la observación X al grupo G_1 si

$$P(Z > t|X) \geq P(Z < t|X);$$

es decir, si $P(G_1|X) \geq P(G_2|X)$, y al grupo G_2 en otro caso. De acuerdo con la expresión (6.17) La regla es:

Asignar la observación X al grupo G_1 si $\Phi(\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1 X) \geq 1-\Phi(\boldsymbol{\gamma}_0+\boldsymbol{\gamma}_1 X)$, lo cual equivale a que $\Phi(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X) \geq \frac{1}{2}$. En términos de $\gamma_0 + \gamma_1 X$, la regla puede expresarse como: asignar X al grupo G_1 si

$$\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X \geq 0, \tag{6.18}$$

[4]Díaz (2007, Cap. 2).

y al grupo G_2 en el otro caso (figura (6.5)). Los parámetros $\boldsymbol{\gamma}_0$ y $\boldsymbol{\gamma}_1$ se estiman a través del método de máxima verosimilitud (con soluciones iterativas), empleando una dicotomización del tipo: $\omega = 0$ si $Z \leq t$ y $\omega = 1$ si $Z > t$. No se requiere que X tenga una distribución multinormal, sino que la distribución condicional de Z dado X sea normal. Esto posibilita la inclusión en el vector X de variables aleatorias discretas.

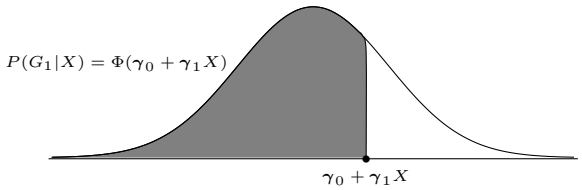

Figura 6.5: Discriminación Probit.

6.5.3 Discriminación con datos multinomiales

La mayoría de los datos procedentes de encuestas corresponden a variables de tipo categórico. Las combinaciones de las categorías constituyen un resultado (valor) de una variable aleatoria multinomial. Por ejemplo, considérense las siguientes cuatro variables categóricas: género (masculino o femenino), credo político (liberal, conservador e independiente), tamaño de la ciudad de residencia (menos de 10.000 habitantes, entre 10.000 y 100.000 y más de 100.000) y nivel de escolaridad (primaria, media, universitaria y de posgrado). El número de posibles valores que toma esta variable multinomial es el producto del número de modalidades de cada una de las variables: $2 \times 3 \times 3 \times 4 = 72$. Para este caso, suponga que se desea predecir si una persona votará en las próximas elecciones, después de habérsele observado alguna de las 72 categorías descritas anteriormente. De esta manera se tienen dos grupos: el grupo G_1 se constituye por los *votantes* y el grupo G_2 por los *no votantes*.

De acuerdo con la regla de Welch, se asigna la observación X a G_1 si

$$\frac{f(X|G_1)}{f(X|G_2)} > \frac{p_2}{p_1}, \tag{6.19}$$

y a G_2 en caso contrario. En el ejemplo, $f(X|G_1)$ se representa por q_{1i}, $i = 1, \ldots, 72$, y $f(X|G_2)$ por q_{2i}, $i = 1, \ldots, 72$, donde q_{1i} es la

probabilidad que una persona del grupo de votantes (G_1) quede en la categoría i; la definición es análoga para q_{2i}. La regla de clasificación (6.19), en términos de las probabilidades multinomiales, es: asignar a la persona identificada con el vector de observaciones X a la población G_1 si

$$\frac{q_{1i}}{q_{2i}} > \frac{p_2}{p_1}, \tag{6.20}$$

y a G_2 en el otro caso. Si las probabilidades q_{1i} y q_{2i} se conocen, se reemplazan en la expresión (6.20) para cada una de las categorías $i = 1, \ldots, 72$; de tal forma que las 72 categorías se particionan en dos clases, una de las cuales se corresponde con individuos del grupo G_1 y la otra con individuos del grupo G_2.

En la práctica, los valores para las probabilidades q_{1i} y q_{2i} no se conocen, y deben estimarse desde los datos muestrales. Suponga que el número de individuos de la i-ésima categoría en los grupos G_1 y G_2 son, respectivamente, n_{1i} y n_{2i}. Se estiman q_{1i} y q_{2i} mediante

$$\widehat{q}_{1i} = \frac{n_{1i}}{N_1} \text{ y} \qquad \widehat{q}_{2i} = \frac{n_{2i}}{N_2}, \tag{6.21}$$

donde $N_1 = \sum_i n_{1i}$ y $N_2 = \sum_i n_{2i}$ son el número de individuos en cada uno de los dos grupos.

Hay situaciones donde las categorías o modalidades de las variables individuales admiten un orden. Si todas las variables tienen categorías ordenadas, entonces se les asigna un rango (puesto) a cada categoría; de esta forma se trabaja de manera directa con los rangos y las reglas usuales de clasificación. Para el caso tratado, el tamaño de la ciudad y el grado de escolaridad son variables de este tipo; por ejemplo, a las categorías de la variable escolaridad se les asignan los números 1, 2, 3 y 4 respectivamente. Se ha demostrado que las funciones de discriminación lineal se desempeñan aceptablemente bien sobre datos ordinales.

Para variables cuyas modalidades no admiten un ordenamiento, por ejemplo el *credo político* de un individuo, el tratamiento debe ser diferente. Así, para una variable con k modalidades no ordenables, estas pueden remplazarse por $(k-1)$ variables "ficticias" (*dummy*) y emplear en estas la discriminación lineal. Para el caso, las tres categorías de la variable *credo político* pueden convertirse en variables ficticias, como se muestra a continuación:

$$Y_1 = \begin{cases} 1, & \text{si es liberal} \\ 0, & \text{en otro caso} \end{cases} \quad Y_2 = \begin{cases} 1, & \text{si es conservador} \\ 0, & \text{en otro caso} \end{cases}$$

Así, el par de variables (Y_1, Y_2) toman los valores $(1, 0)$ para un liberal, $(0, 1)$ para un conservador y $(0, 0)$ para un independiente.

6.5.4 Clasificación mediante la técnica de "el vecino más cercano"

El método de clasificación llamado *"el vecino más cercano"* se considera una técnica de tipo no paramétrico. Para el procedimiento se determina la distancia de Mahalanobis de una observación X_i respecto a las demás observaciones X_j, mediante

$$D_{ij} = (X_i - X_j)' \boldsymbol{S}_p^{-1} (X_i - X_j), \quad i \neq j. \tag{6.22}$$

Para clasificar la observación X_i en uno de dos grupos, se examinan los k puntos más cercanos a X_i. Si la mayoría de estos k puntos pertenecen al grupo G_1, se asigna la observación X_i a G_1, en otro caso, se asigna a G_2. Si se nota el número de individuos (objetos) de G_1 por k_1 y a los restantes por k_2 en G_2, con $k = k_1 + k_2$, entonces la regla se expresa también como: asignar X_i a G_1 si

$$k_1 > k_2, \tag{6.23}$$

y G_2 en otro caso. Si los tamaños muestrales de cada grupo son n_1 y n_2 respectivamente, la decisión es: asignar X_i a G_1 si

$$\frac{k_1}{n_1} > \frac{k_2}{n_2}. \tag{6.24}$$

De manera coloquial, una observación X_i se asigna al grupo donde se "inclinen" la mayoría de sus vecinos.

Además, si se consideran las probabilidades a priori: asignar x_i a G_1 si

$$\frac{k_1/n_1}{k_2/n_2} > \frac{p_2}{p_1}. \tag{6.25}$$

Estas reglas se pueden extender a más de dos grupos. Así, en (6.24): se asigna la observación al grupo que tenga la más alta proporción k_j/n_j, donde k_j es el número de observaciones en el grupo G_j entre las k observaciones más cercanas a X_i.

Respecto al valor k, se sugiere tomar un valor cercano a $\sqrt{n_i}$ para algún n_i típico. En la práctica, se puede ensayar con varios valores de k y usar el que menor tasa de error provoque.

6.6 Selección de variables

La selección de variables en el análisis discriminante está asociada al uso que se pretenda dar a la metodología. De acuerdo con los dos objetivos presentados al comienzo de este capítulo, uno corresponde a la *separación* de grupos y el otro a la *localización o clasificación* de observaciones o casos. Las metodologías empleadas para la separación de grupos se relacionan con las estadísticas parciales T^2 o Lambda de Wilks (Λ), con las cuales se verifica la influencia de un subconjunto de variables en la separación (diferencia de medias) de dos o más grupos. En esta parte se comentan algunas metodologías para el segundo propósito (Díaz, 1999: Cap. 3).

Es importante advertir sobre el cuidado que se debe tener al intercambiar el uso de metodologías cuyos propósitos son la separación de grupos o la localización de observaciones, respectivamente.

El problema sobre la contribución de cada variable en la discriminación, como se procede en el análisis de regresión, está ligado a la búsqueda de la función de predicción con las variables que mejor contribuyan a la discriminación. Naturalmente, se procura incorporar al modelo el menor número variables predictoras (principio de parsimonia). Uno de los criterios de selección de variables es escoger el subconjunto que produzca la menor tasa de error.

A continuación se comentan los procedimientos más empleados, los cuales están incorporados en la mayoría de los paquetes estadísticos.

Para el caso de dos grupos se recomiendan dos procedimientos:

1. Las estadísticas F parciales con niveles de significancia nominal entre 0.10 y 0.25. Con estas estadísticas, se observa el aporte "extra" que cada variable hace al modelo; una vez que han ingresado las demás, se incorporan las que tengan el mayor valor F.

2. Un estimador de la probabilidad de clasificación correcta basado en la distancia de Mahalanobis entre dos grupos (McLachlan 1992, 366–367).

Un mecanismo formal para la selección del "mejor" subconjunto de variables en cualquier problema de modelamiento requiere un criterio que evalúe la bondad del ajuste, un procedimiento para el cálculo (generalmente computacional) y, tal vez, una regla necesaria para "frenar" el

proceso (Krzanowski 1995, pág. 41). Entre de los procedimientos para el cálculo de la bondad del ajuste en la selección de variables se cuentan la selección *hacia adelante* (*forward*), la eliminación hacia *atrás* (*backward*) y la selección *"stepwise" (selección paso a paso).*

En la selección hacia adelante (*forward*) la función de clasificación se inicia con la variable que bajo algún criterio sea la más apropiada (generalmente a través de la estadística F). En la segunda etapa se adiciona, entre las restantes $(p-1)$ variables, la que mejor desempeño muestre en la regla de clasificación, luego se agrega a estas dos variables una entre las $(p-2)$ restantes de mejor desempeño, y así sucesivamente.

La eliminación hacia atrás (*backward*) trabaja en sentido opuesto a la técnica anterior. Se empieza la función con todas las p variables, se remueve en cada etapa la variable que menos afecte el "buen desempeño" de la función de clasificación.

La estrategia de selección basada en el método "*stepwise* " funciona en forma parecida al procedimiento de selección hacia adelante, la diferencia es que, en cada etapa, una de las variables ya incorporadas al modelo puede ser removida sin que menoscabe el desempeño de la función de clasificación.

La tres estrategias anteriores requieren una regla para finalizar el proceso en términos de mejoramiento o deterioro. La regla natural es terminar el proceso cuando la adición de nuevas variables no incremente significativamente el buen desempeño de la función o cuando la exclusión de cualquiera de las variables ya incorporadas al modelo deteriore su desempeño. El término "desempeño" puede ser juzgado a través de la tasa de clasificación, de la estadística Lambda de Wilks (Λ) para un subconjunto de variables o de algún incremento en términos de suma de cuadrados, como se hace en análisis de regresión.

Otro procedimiento consiste en combinar el *stepwise* con el criterio de estimación del error mediante validación cruzada. En este procedimiento se excluye cada observación, se selecciona un subconjunto de variables para construir la regla de clasificación, y luego se clasifica la observación excluida empleando reglas de clasificación lineal computadas con las variables seleccionadas. Las tasas de error resultantes son usadas para escoger la variable que debe incorporarse al modelo en cada etapa.

6.7 Procesamiento de datos con R

En esta sección se llevan a cabo los cálculos correspondientes al ejemplo tratado en la sección 6.2 y se desarrolla un ejemplo para ilustrar cómo realizar análisis discriminante mediante R.

Cálculos para la sección 6.2

Introducción de los vectores $\widehat{X}_1$, $\widehat{X}_2$ y de la matriz $\boldsymbol{S}$

```
# vector de medias 1
Xb1<-matrix(c(12.57,9.57,11.49,7.97))
# vector de medias 2
Xb2<-matrix(c(8.75,5.33,8.5,4.75))
# matriz S
S<-matrix(c(11.2553,9.4042,7.1489,3.3830,
            9.4042,13.5318,7.3830,2.5532,
            7.1489,7.3830,11.5744,2.6170,
            3.3830,2.5532,2.6170,5.8085),
            nrow=4)
```

Los vectores $\widehat{\boldsymbol{b}}'$ y $\overline{X}'_c$ se obtienen de la siguiente forma:

```
#b'
bpg<-t(Xb1-Xb2)%*%solve(S)
# X_c
Xc<-1/2*(Xb1+Xb2)
b'X_c
bpg%*%Xc
```

Ejemplo sobre discriminación para varios grupos

El siguiente ejemplo, tomado de Johnson (2000: 235–243) y citado por Díaz (2007: 329–333), se usa para ilustrar cómo se hace análisis discriminante en R.

Se quiere encontrar una regla para discriminar entre cuatro grupos de semillas de trigo. Los grupos se definen de acuerdo con el sitio de cultivo

y con la variedad del trigo. Así, los grupos 1 y 2 se corresponden con dos variedades (ARKAN y ARTHUR) cultivadas en un primer sitio (MAS0), mientras que los grupos 3 y 4 se corresponden con las mismas variedades cultivadas en un segundo sitio (VLAD12).

La investigación apunta a encontrar una manera de identificar las semillas de trigo con base en medidas físicas como área, perímetro, longitud y ancho del grano.

Cada grano tiene un pliegue, de manera que se optó por tomar medidas tanto con el pliegue a derecha como con el pliegue hacia abajo. Las variables que se midieron en el grano, cuando el pliegue estaba hacia abajo, son: raíz cuadrada del área (`DA`), perímetro (`DP`), longitud (`DL`) y ancho (`DB`). Las variables `RA`, `RP`, `RL` y `RB` se definen de manera análoga, excepto que el grano se midió con el pliegue a la derecha.

A continuación se hace la lectura de los datos, los cuales se encuentran en el directorio de trabajo, en un de texto plano.

```
# lectura de los datos
ejemp<-read.table("ejemplo8_2.txt",header=TRUE)
# transformación mediante raíz cuadrada
ejemp$DA<-sqrt(ejemp$DA)
ejemp$RA<-sqrt(ejemp$RA)
# definición del factor que identifica el grupo
ejemp$GRP<-factor(ejemp$GRP)
head(ejemp)
```

Después de ejecutar las órdenes anteriores, se obtiene la siguiente salida que corresponde a las 6 primeras filas del archivo de datos.

```
  DA  DP  DL DB  RA    RP RL RB GRP
54.45 219 89 43 56.60 226 89 47  1
55.15 221 91 46 56.26 224 91 46  1
53.92 223 90 44 55.09 223 91 44  1
52.23 212 87 41 53.54 215 88 44  1
51.56 207 78 42 52.98 211 81 44  1
50.43 203 82 41 51.22 207 82 42  1
```

El siguiente paso es verificar si se puede asumir la igualdad de las matrices de varianza–covarianza de los grupos. Para saberlo, se lleva a cabo el contraste de la hipótesis $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{\Sigma_3} = \mathbf{\Sigma_4}$. La prueba arroja un valor $F = 4.228$ y un $p-$valor de 3.849927×10^{-44} con lo que se concluye que no se puede asumir homogeneidad de las matrices de varianzas y covarianzas[5]. Se discutió en la sección 6.3.2 que, en el caso de grupos con matrices de covarianzas distintas, la función de discriminación cuadrática clasifica mejor que la lineal; en consecuencia se construye la regla de discriminación con la función `qda()` como sigue a continuación:

```
library(MASS)
zq<-qda(GRP ~.,ejemp)
```

La función `qda()` usa como probabilidades a priori n_i/N. En caso de que se tengan otros valores para estas probabilidades, pueden entregarse mediante la opción `prior=c(p1,p2,p3,p4)` teniendo en cuenta que los valores p_i son mayores que cero y menores que uno y su suma debe ser uno.

Suponga que se tiene una nueva observación y deseamos usar la regla construida para asignarla a uno de los cuatro grupos. La nueva observación es

```
nuevo<-data.frame(DA=54.36,DP=220.07,DL=90.08,
                  DB=44.84,RA=53.86,
                  RP=220.17, RL=90.27,
                  RB=43.53)
```

y la aplicación de la regla se hace mediante

[5]El lector puede consultar en Díaz (2007, Cap. 4) la teoría relacionada con esta prueba y una función en R para realizarla.

```
predict(zq,nuevo)
```

La salida de la orden anterior es:

```
$class
[1] 4
Levels: 1 2 3 4

$posterior
          1          2         3         4
1 0.1895238 0.02623314 0.3420116 0.4422315
```

se concluye que la regla clasifica a la nueva observación en el grupo 4. La última linea corresponde a las probabilidades a posteriori de acuerdo con la ecuación (6.8).

Para la estimación de la tasa de error, como se trató en la sección 6.4.1, se aplica la regla de discriminación a cada una de las observaciones que se usaron para construirla y se observa si es clasificada correctamente o no. El código y la salida (las 6 primeras filas) son:

```
clasifq<-predict(zq,ejemp[,-9])$class
head(data.frame(Grupo=ejemp$GRP,Casificada=clasifq))
```

```
  Grupo Casificada
1     1          1
2     1          3
3     1          3
4     1          1
5     1          1
6     1          1
```

La primera observación pertenece al grupo uno y fue asignada correctamente, mientras que las observaciones 2 y 3 fueron asignadas erróneamente al grupo 3. Con el siguiente código se obtiene una tabla que permite calcular las tasas de error de clasificación.

```
tabla<-table(ejemp$GRP,clasifq)
addmargins(tabla)
     clasifq
```

```
       1   2   3   4 Sum
1     24   1   9   2  36
2      2   8   2  24  36
3      5   1  41   3  50
4      0   2   1  47  50
Sum   31  12  53  76 172
```

En la salida anterior se observa que en el grupo 1 hay 36 observaciones, de las cuales 24 son asignadas correctamente, una es asignada erróneamente al grupo 2, 9 al 3 y 2 al grupo 4. La tasa de error aparente se calcula mediante el código

```
(sum(tabla)-sum(diag(tabla)) )/ sum(tabla)
0.3023256
```

Para el análisis discriminante lineal se usa la función `lda()`. El código es igual; solo debe escribirse `lda()` en lugar de `qda()`.

6.8 Procedimiento DISCRIM del paquete SAS

Para un conjunto de observaciones que contienen variables cuantitativas y una variable de clasificación, que define el grupo de cada observación, el PROC DISCRIM desarrolla un criterio de discriminación para asignar cada observación en uno de los grupos. SAS tiene el procedimiento STEPDISC, el cual desarrolla análisis discriminante con selección de variables (tipo *stepwise*, *forward* y *backward*).

Al frente (o debajo) de cada instrucción se explica su propósito dentro de los símbolos $/*$ y $*/$.

```
/* Análisis discriminante*/
DATA nombre SAS;/*nombre del archivo de datos*/
/*variables, incluyendo la de clasificación*/
INPUT variables;
CARDS;          /*ingreso de datos*/

escribir aquí los datos
```

```
;
PROC DISCRIM CROSSVALIDATE POOL=YES CROSSLIST;\\
  /*desarrolla discriminación asumiendo igualdad
   de las matrices de covarianzas e imprime la
   validación cruzada por observación*/
CLASS variable; /*se indica la variable que
                  define los  grupos*/
/*se escriben las variables cuantitativas para
   el análisis*/
VAR lista de variables;
PRIORS EQUAL$|$PROP$|$probabilidades;\\
/*(EQUAL) toma iguales las probabilidades
   a priori para cada grupo*/
/*(PROP) hace las probabilidades proporcionales
  a los tamaños de grupo*/
/*también se puede dar las probabilidades a priori
  para cada grupo*/
/*Ejemplo, para tres grupos 1, 2 y 3, se escribe
  PRIORS '1'=0.25 '2'=0.35 '3'=0.40;*/
/*Por defecto se considera la opción EQUAL  */
 RUN;
```

6.9 Ejercicios

1. El marco de datos `biopsy` de la librería `MASS` contiene información sobre 699 biopsias de tumores de mama que han sido clasificados como benignos o malignos. Las nueve variables de cada biopsia son una clasificación o valoración (0 a 10) efectuada por el coordinador médico. Analice los datos. En particular, investigue las diferencias en los dos tipos de tumor, encuentre una regla para clasificar los tumores basada únicamente en las variables de la biopsia y evaluar la precisión de la regla.

2. Se realizó un ensayo clínico para determinar la capacidad analgésica de la acupuntura en el tratamiento de los dolores lumbares. Para este fin se aplicó el tratamiento a 40 pacientes, colocando las agujas en los lugares indicados por la medicina tradicional china, a otros 40 se aplicó un tratamiento placebo consistente en situar

Tabla 6.2: Datos de acupuntura.

Y	trat	sexo	edad	psic	Y	trat	sexo	edad	psic	Y	trat	sexo	edad	psic
1	0	0	49	1	1	1	1	34	2	0	0	1	65	0
0	0	1	23	0	0	1	1	53	1	0	0	1	53	1
0	0	1	43	1	0	1	0	35	0	0	0	0	44	2
0	0	0	25	0	0	1	0	32	1	0	0	0	37	2
1	0	0	22	3	0	1	0	23	0	0	0	0	42	0
0	0	0	34	2	0	1	1	32	1	0	0	1	59	0
1	0	1	55	2	1	1	1	32	2	0	1	1	36	1
1	0	1	35	2	1	1	0	24	3	0	1	0	18	0
0	0	1	43	0	1	1	0	35	4	0	1	0	44	0
0	0	1	35	1	1	1	1	47	2	1	1	0	53	4
0	0	0	51	0	1	1	1	38	2	1	1	1	43	3
0	0	0	41	3	0	1	1	23	0	1	1	1	45	4
0	0	0	41	4	1	1	0	43	4	1	1	0	44	3
0	0	0	63	0	0	0	0	35	0	0	1	1	24	1
0	0	1	23	0	0	0	0	54	0	0	1	1	47	0
0	0	0	65	3	0	0	1	63	2	0	1	0	45	2
0	0	0	45	0	0	0	1	41	0	1	1	1	34	3
0	0	0	28	0	1	0	0	33	2	1	1	0	52	4
0	0	0	37	1	0	0	1	31	0	0	1	1	23	0
0	0	1	32	0	1	0	0	37	3	0	1	1	26	1
0	1	1	45	1	0	0	0	52	1	1	1	1	26	3
0	1	1	43	1	0	0	1	43	0	0	1	1	53	1
0	1	1	34	2	0	0	0	35	0	0	1	1	23	1
0	1	0	35	0	0	0	0	43	0	0	1	1	24	0
0	1	0	42	1	0	0	0	64	0	0	1	0	24	1
1	1	1	40	0	0	0	1	65	1	0	1	0	33	1
1	1	1	54	2	0	0	1	29	1					

las agujas de forma aleatoria sobre la superficie cutánea (tradicional=1, placebo=0). Además, se sabe de cada uno de ellos el sexo (mujer=0, hombre=1), la edad y el historial de consumo de psicofármacos (no consumió=0, consumo ocasional=1, regular=2, frecuente=3, dependencia=4). Transcurridos dos meses de tratamiento, el individuo declara que ha experimentado mejoría

($Y = 1$) o que los dolores se mantienen o que han aumentado ($Y = 0$). Los datos se muestran en la tabla 6.2. Obtenga una regla de discriminación a partir de un modelo logístico que permita clasificar a nuevos pacientes en uno de los dos grupos. ¿Qué variables deben retenerse en el modelo?

Capítulo 7

Métodos no paramétricos

7.1 Introducción

Las técnicas para el análisis de datos procedentes de experimentos, en su mayoría, suponen que estos siguen un determinado modelo distribucional, generalmente el *normal*, y una escala de medición; de manera que la técnica resulta válida cuando se satisface el modelo probabilístico y la escala de medida asociada a las variables en consideración.

Para el caso de datos categóricos, cuya escala de medición es ordinal o nominal, el cumplimiento de tales requerimientos es aún más difícil. Cuando no se cumple o se tienen serias dudas sobre el modelo supuesto, se han desarrollado algunos procedimientos alternativos:

- El primer procedimiento consiste en *transformar* adecuadamente los datos para que se "comporten" de acuerdo con el modelo probabilístico requerido por el procedimiento; tal es el caso de la transformación de *Box-Cox* o de algunos casos especiales, como la familia de transformaciones de Tukey, la transformación raíz cuadrada, la transformación inversa, la transformación logarítmica, entre otras.

- La segunda salida es justificar el *Teorema Central del Límite*, el cual, en su forma más amplia, garantiza que para muestras *suficientemente grandes* extraídas de cualquier población, que tenga media y varianza finitas, la media muestral ($\overline{X}$) tiene aproxima-

damente una distribución normal. En estas condiciones, se puede aplicar el teorema para la construcción de algunos intervalos de confianza y la verificación de hipótesis sobre la media poblacional.

- Como tercera opción está la estadística *robusta*, cuyos procedimientos son resistentes o insensibles a desviaciones de los datos respecto al modelo supuesto.

- Finalmente están las técnicas *libres de distribución* o *no paramétricas*, las cuales se basan en una función de las observaciones muestrales, que es una variable aleatoria cuya distribución no depende de la distribución específica de la población de donde se extrajo la muestra. De otra manera, una prueba estadística no paramétrica es aquella cuyo modelo no especifica las condiciones de los parámetros de la población de donde se extrajo la muestra, salvo algunas pocas suposiciones y mucho más débiles que las requeridas en el caso paramétrico.

Los procedimientos estadísticos clásicos se basan, usualmente, en la distribución normal. Los datos del mundo real no siempre se ajustan a una distribución normal, o como dice el aforismo estadístico, "*lo más anormal es la normalidad*". En consecuencia, la aplicación de la teoría normal puede dejar muchas insatisfacciones por intentar "disfrazar" la realidad.

En la primera parte de este capítulo se presentan las técnicas no paramétricas para una muestra. En la segunda parte se examinan las pruebas de localización para dos muestras. En la tercera y cuarta parte se desarrolla el análisis de diseños experimentales completamente al azar y en bloques completos.

7.2 Pruebas de localización en una muestra

En problemas de una muestra se considera un conjunto de datos extraídos de una única población, con los cuales se pretende hacer inferencia. Se asume, cuando sea necesario, que la población tiene una distribución continua y simétrica. La inferencia estadística se centra en

un parámetro de localización (la media o la mediana, por ejemplo) con un enfoque no paramétrico análogo a la prueba *normal* (varianza conocida) o a la prueba *t-Student* (varianza desconocida) para la hipótesis $H_0 : \mu = \mu_0$ o también para el caso $H_0 : \mu_X - \mu_Y = \mu_D = 0$, en muestras pareadas.

La prueba *t-Student* solamente debe aplicarse cuando se reúnen las siguientes condiciones:

1. Los datos son independientes (muestra aleatoria).
2. Los datos están en escala al menos de intervalo.
3. La población debe ser (aproximadamente) normalmente distribuida.
4. La varianza deben ser constante.

En los procedimientos no paramétricos simplemente se requiere la condición (1), es decir, tan solo se debe garantizar que los datos sean obtenidos de la misma población de manera aleatoria e independiente entre sí. Este hecho se sostiene desde la estructura misma de planeamiento y desarrollo de un estudio.

A continuación se enuncian algunas ventajas de las llamadas pruebas no paramétricas:

1. Son menos exigentes que los métodos paramétricos.
2. Las probabilidades para la mayoría de las estadísticas de prueba no paramétrica son en general exactas, excepto cuando se usan aproximaciones asintóticas.
3. Son independientes de la forma poblacional de donde se extrae la muestra.
4. Son más fáciles de aplicar, por la simplicidad de sus cálculos.
5. Permiten trabajar con datos de diferentes poblaciones; situación que no es posible en el caso paramétrico.
6. Pueden ser aplicadas en datos que no posean escala de intervalo;

7. Son más eficientes que las técnicas paramétricas cuando los datos de la población no tienen distribución normal. Cuando la población es normal, su eficiencia es levemente inferior a su competencia.

Como no todo siempre es bueno, las siguientes son algunas de las restricciones de las técnicas no paramétricas:

1. En general no consideran la verdadera magnitud de los datos.
2. Cuando se satisfacen los requerimientos del modelo estadístico, las pruebas paramétricas son más potentes.
3. En general, no permiten probar interacciones, excepto en condiciones especiales de aditividad.
4. La obtención, utilización e interpretación de las tablas en general son más complejas.

Al suponer únicamente una población con distribución continua, se requiere una medida de localización para tal distribución. Como no hay supuestos respecto a la forma de la distribución, los parámetros usuales de localización, *la media y la mediana*, no necesariamente son iguales. Una primera ventaja de la mediana sobre la media es que esta siempre existe. Otra ventaja de la mediana es su *resistencia (robustez)* a datos extremos o *outliers* en la distribución; es decir, datos extremos ejercerán una pequeña influencia sobre la mediana, mientras que su influencia sobre la media es alta. Se usará, por ahora, la mediana θ como un parámetro para localizar la distribución.

7.2.1 Prueba del signo

Se presenta una de las más simples y antiguas pruebas no paramétricas; la *prueba del signo*. Se asume que la *mediana* es un valor único en cualquier distribución; luego por definición de mediana, esta es un valor que divide la población en dos subpoblaciones de igual frecuencia (del 50%), es decir, la mediana θ es tal que $P(X > \theta_X) = P(X < \theta_X) = 0.5$.

Para verificar la hipótesis

$$H_0 : \ \theta_X = \theta_0, \quad \text{frente a} \quad H_1 : \ \theta_X > \theta_0 \tag{7.1}$$

se extrae una muestra de n observaciones $X_1, \cdots, X_n$ de la población F_X con mediana θ desconocida. Las otras dos hipótesis alternativas se trabajan en forma similar.

Si se transforman los datos de la muestra anterior desplazándolos una cantidad θ_0 (centrar en torno a 0),

$$Y_1 = X_1 - \theta_0, Y_2 = X_2 - \theta_0, \cdots, \ Y_n = X_n - \theta_0,$$

entonces verificar la hipótesis (7.1) es equivalente a verificar la hipótesis de que la mediana θ_Y de los "nuevos" datos es cero, es decir,

$$H_0 : \ \theta_Y = 0, \text{ frente a } H_1 : \ \theta_Y > 0. \tag{7.2}$$

Si los datos muestrales son consistentes con la hipótesis nula (7.1), aproximadamente la mitad de ellos "caerán" por debajo del valor θ_0 y la otra mitad por encima. Para los datos transformados a Y_i, la hipótesis nula equivale a que el número de datos positivos es igual (aproximadamente) al de datos negativos. Se nota que la hipótesis nula dicotomiza las observaciones muestrales: por una parte las que están por debajo de la mediana; por otra, las que están por encima (positivas y negativas para las Y_i).

Se define la estadística

$$\mathbf{S} = \sum_{i=1}^{n} s(Y_i), \quad i = 1, \cdots, n; \text{ donde } \ s(x) = \begin{cases} 1, & \text{si } x > 0 \\ 0, & \text{si } x \leq 0 \end{cases} \tag{7.3}$$

que cuenta "el número de observaciones Y_i positivas" o, equivalentemente, el número de observaciones X_i por encima de θ_0.

Una *región de rechazo* apropiada, de acuerdo con la hipótesis alternativa, es determinada por aquel valor k_α tal que $P_{H_0}(\mathbf{S} \geq k_\alpha) = \alpha$, donde α es el nivel de significancia de la prueba. La región de rechazo $\mathbf{C}$, para un nivel de significancia α, es

$$\mathbf{C} = \{k \mid k \geq k_\alpha\} \tag{7.4}$$

La escritura P_{H_0} indica que se debe encontrar la distribución de $\mathbf{S}$ bajo la hipótesis nula. Bajo la hipótesis nula (7.2), $s(X_1), s(X_2), \cdots, s(X_n)$ son variables aleatorias independientes e idénticamente distribuidas (*iid*), cada una binomial (Bernoulli) con parámetros $n = 1$ y $p = P(X > 0) = 1 - F(0) = \frac{1}{2}$; se nota $Y_i \sim B(1, \frac{1}{2})$, para $i = 1, \cdots, n$. Puesto que $\mathbf{S}$ es

la suma de variables aleatorias *iid* tipo $B(1, \frac{1}{2})$, esta tiene distribución $B(n, \frac{1}{2})$.

Los valores de k_α son los enteros más pequeños tal que

$$\sum_{k=k_\alpha}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n \leq \alpha \tag{7.5}$$

Similarmente, para alternativas del tipo $H_1 : \ \theta_x < \theta_0$, la región crítica es definida por el valor k^*_α tal que

$$\sum_{k=0}^{k^*_\alpha} \binom{n}{k} \left(\frac{1}{2}\right)^n \leq \alpha \tag{7.6}$$

Para alternativas bilaterales de la forma $H_1 : \ \theta_X \neq \theta_0$ (o $\theta_Y \neq 0$), la región de rechazo está conformada por los k tales que

$$\sum_{k=0}^{k^*_{\alpha/2}} \binom{n}{k} \left(\frac{1}{2}\right)^n \leq \alpha \ \text{ y } \ \sum_{k=k^*_{\alpha/2}}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n \leq \alpha \tag{7.7}$$

Esto permite encontrar los valores críticos k_α, k^*_α, $k_{\alpha/2}$ o $k^*_{\alpha/2}$ desde las tablas para la distribución binomial, respectivamente.

En estudio sobre pacientes (de cierto perfil) con problemas cardiacos se observaron los tiempos entre dos preinfartos consecutivos[1]. Se tomó una muestra de los tiempos entre los dos eventos (en semanas) durante un periodo; los resultados son los siguientes:

Pac.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Tpo.	7.1	1.2	1.9	2.5	6.1	0.2	1.2	0.2	1.1	1.6	0.3	10.4	3.1	13.3	8.1	5.3

Con esta información se quiere verificar la hipótesis que la mediana de los tiempos es 5, es decir $H_0 : \ \theta = 5$, frente a $H_1 : \ \theta < 5$. Si a cada uno de los datos se le resta 5, los datos resultantes son

$Y_i = X_i - 5$:	2.1	−3.8	−3.1	−2.5	1.1	−4.8	−3.8	−4.8	−3.9	−3.4	−4.7
	5.4	−1.9	8.3	3.1	0.3						

[1]Estos datos tienen una distribución tipo exponencial con parámetro θ.

El valor de la estadística **S** es

$$\mathbf{S} = \sum_{i=1}^{16} s(X_i - 5) = 6$$

De la ecuación (7.6), y por la tabla A.1, se obtiene

$$P(Y \leq 6) = \sum_{k=0}^{6} \binom{16}{k} \left(\frac{1}{2}\right)^{16} = 0.2272$$

Se concluye que, incluso con un valor de $\alpha = 0.10$, no se rechaza la hipótesis nula de que la mediana del tiempo entre dos preinfartos consecutivos es al menos 5.

7.2.2 Muestras pareadas

Suponga que se tienen dos muestras, las cuales no son independientes porque existe un apareamiento natural entre la observación X_i de la primera muestra con la observación Y_i de la segunda muestra para todo i. Por ejemplo, cuando se aplica un tratamiento a un individuo y se observa su respuesta "pre" (X) y su respuesta "pos" (Y) al tratamiento; otra situación es cuando los objetos son mezclados de acuerdo con algún criterio de homogeneidad. Otro caso son los individuos con un mismo cociente intelectual (CI) o con los mismos rasgos familiares. Con tales pares, el procedimiento es frecuentemente referido como *observaciones pareadas* o *pares mezclados*.

Se trata de hacer inferencia sobre una población de la cual se ha extraído una muestra de la forma $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$. Se puede considerar también cada pareja como un bloque en el que las unidades experimentales son homogéneas respecto a un atributo. Se quiere entonces detectar la existencia de posibles diferencias entre las dos mediciones.

Con la muestra aleatoria de los n pares se pueden formar las diferencias muestrales $D_i = X_i - Y_i$, para $i = 1, \cdots, n$. Para verificar la hipótesis $H_0 : \theta_D = 0$, se procede en forma similar al caso anterior; basta cambiar X_i por D_i.

Cabe precisar que, en general, la *mediana de las diferencias no necesariamente es la diferencia de las medianas*, es decir θ_D no siempre es igual a la diferencia entre θ_X y θ_Y; en cambio para la media si lo es. Para que

la mediana de las diferencias sea igual a la diferencia de las medianas es necesario y suficiente que las distribuciones de las variables aleatorias (marginales) y la distribución de la diferencia sean simétricas.

Suponga que se quiere comprobar la eficiencia de una nueva dieta. Se seleccionaron nueve individuos aleatoriamente y se registró su peso antes de someterse al tratamiento. Sus pesos antes y después del programa aparecen en seguida:

Individuo	1	2	3	4	5	6	7	8	9
Peso antes:	80	82	75	90	98	87	100	107	103
Peso después:	79	83	75	81	95	86	101	105	100
$\text{sig}(X_i - Y_i)$:	+	-	0	+	+	+	-	+	+

La hipótesis nula por verificar es

$$H_0 : \ \theta = P(X > Y) = PX < Y) = \frac{1}{2}$$
$$\text{frente a} \quad H_1 : \ P(X > Y) > \frac{1}{2} \text{ o } \theta = P(X < Y) < \frac{1}{2}$$

El valor de **S** es 6. Como hay un empate en la observación 3, se descarta esta observación. La probabilidad bajo H_0 de obtener un valor menor que o igual a 2 es

$$P(Y \leq 2) = \sum_{k=0}^{2} \binom{8}{k} \left(\frac{1}{2}\right)^8 = 0.1446.$$

Con un nivel de significancia $\alpha = 0.10$ se concluirá que no hay evidencia para rechazar la hipótesis H_0. Es decir, con estos datos, se puede afirmar que la dieta no influye en el cambio de peso.

Rangos

Dado un conjunto de observaciones, el rango de una observación es el "puesto" que ocupa la observación respecto a las demás.

Formalmente, dada la muestra aleatoria $X_1, \cdots, X_n$, de una población con función de distribución $F_X(x)$ continua, se define el rango de X_i

entre $X_1, \cdots, X_n$ mediante

$$R(X_i) = R_i = \sum_{j=1}^{n} s(X_i - X_j), \quad i = 1, \cdots, n, \tag{7.8}$$

con $s(x)$ definido por (7.3).

Se advierte que la notación R_i significa el rango de la observación X_i, donde $1 \leq n_i \leq n$. Por la continuidad asumida, los *empates* entre $X_1, \cdots, X_n$ se presentan con probabilidad cero; en consecuencia, pueden descartarse; así, $(R_1, R_2, \cdots, R_n)$ es alguna permutación de los números $\{1, 2, \cdots, n\}$.

A manera de ilustración: sean las observaciones 4.0, 8.6 3.2 2.0 1.9 0.5 5.8. En este caso $X_1 = 4.0$, $X_2 = 8.6$, $X_3 = 3.2$, $X_4 = 2.0$, $X_5 = 1.9$, $X_6 = 0.5$ y $X_7 = 5.8$, el rango de $X_1 = 4.0$ es $R(X_1) = R_1 = 5$, y sucesivamente. El conjunto de observaciones X_i se ha transformado en los R_i: 5, 7, 4, 3, 2, 1 y 6, respectivamente.

Aunque la ocurrencia de empates está descartada en teoría, en la práctica aparecen observaciones empatadas o iguales; esto puede deberse al muestreo mismo. El método más empleado para manejar empates consiste en asignar a cada observación del grupo empatado la media aritmética de los rangos que le corresponderían si fuesen diferentes. Aunque la media de los rangos no se afecta, su varianza sí disminuye; razón por la que algunas pruebas introducen correcciones por empates.

7.2.3 Prueba de rango signado de Wilcoxon

La prueba del signo utiliza únicamente los signos de las diferencias entre cada observación y la mediana asumida en la hipótesis nula; el contraste ignora la distancia entre la observación y la mediana θ_0, es decir, no considera la magnitud de las observaciones. Una prueba que además del orden de las observaciones considera o tiene en cuenta su magnitud fue propuesta por Wilcoxon (1945); se conoce como la *prueba de rango signado de Wilcoxon*. Es una alternativa para pruebas de localización, la cual es sensible a la magnitud y al signo de las diferencias de las observaciones respecto a la mediana supuesta en H_0.

Considere una muestra aleatoria $X_1, X_2, \cdots, X_n$ extraída de una población *continua y simétrica* alrededor de su mediana θ. Bajo la hipótesis

nula,

$$H_0 : \ \theta = \theta_0 \tag{7.9}$$

la diferencia $D_i = X_i - \theta_0$ es simétrica alrededor de cero; por tanto, las diferencias positivas y negativas son iguales en valor absoluto y equiprobables, es decir,

$$P(D_i \leq -k) = P(D_i \geq k) = 1 - P(D_i \leq k).$$

La hipótesis (7.9) es equivalente a la hipótesis

$$H_0 : \ \theta_D = 0 \tag{7.10}$$

Bajo el supuesto de continuidad no habrá problemas, al menos teóricamente, con diferencias nulas o empatadas en valor absoluto. A las diferencias absolutas $|D_1|, |D_2|, \cdots, |D_n|$ se les asigna el rango respectivo $R(|D_i|) = R_i$, entonces

$$T^+ = \sum_{i=1}^{n} R_i s(D_i) = \sum_{i=1}^{n} i s(D_i) \tag{7.11}$$

es la estadística de Wilcoxon para verificar $H_0 : \ \theta_D = 0$. La estadística T^+ resalta la consideración de las observaciones a la derecha de la mediana supuesta; de ahí el calificativo de *rango signado*. También se puede emplear la estadística T^-, la cual se calcula sobre las diferencias negativas; además, $T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.

Cuando la distribución tiene mediana $\theta > 0$ y es sesgada a la derecha, por ejemplo, la distribución se traslada (restando θ); así las diferencias positivas tenderán a estar más lejos de 0 que las negativas. En este caso, T^+ tenderá a ser grande y provocará un rechazo de $H_0 : \ \theta_D = 0$.

La figura (7.1) muestra que T^+ tiende a ser grande aun cuando la mediana sea 0. Por tanto, el supuesto de simetría es necesario para evitar una interpretación ambigua de valores grandes de T^+. De manera que si la mediana de la población es conocida, la prueba de Wilcoxon se convierte en una prueba de simetría.

En la tabla 7.1 se insinúa cómo construir su distribución en una muestra de tamaño 4. Se muestran las $2^4 = 16$ posibles asignaciones de signos a las cuatro diferencias. La asignación por observación tiene una probabilidad de 1/2, luego la probabilidad de una asignación completa a las cuatro observaciones es $\frac{1}{2^4} = \frac{1}{16}$.

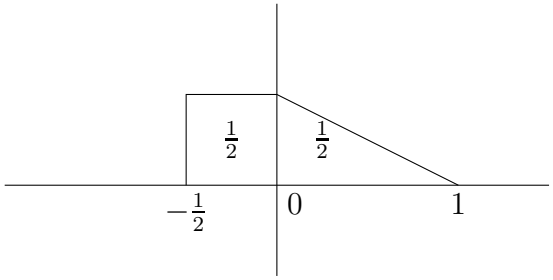

Figura 7.1: Distribución sesgada de mediana 0.

Tabla 7.1: Distribución de T^+ con $n = 4$.

	Rangos					
	1	2	3	4	T^+	$P(T^+)$
Signos						
	+	+	+	+	10	1/16
	−	+	+	+	9	1/16
	+	−	+	+	8	1/16
	+	+	−	+	7	
	−	−	+	+	7	2/16
	+	+	+	−	6	
	−	+	−	+	6	2/16
	+	−	−	+	5	
	−	+	+	−	5	2/16
	+	−	+	−	4	
	−	−	−	+	4	2/16
	+	+	−	−	3	
	−	−	+	−	3	2/16
	−	+	−	+	2	1/16
	+	−	−	−	1	1/16
	−	−	−	−	0	1/16

La tabla 7.1 muestra la simetría de la distribución. Para rechazar $H_0 : \theta_D = 0$, en pruebas de dos colas, se puede tomar $\alpha = P(T^+ \geq 10) + P(T^+ \leq 0) = 1/16 + 1/16 = 0.125$ como tamaño mínimo de la prueba. Para pruebas unilaterales, el mínimo valor de α se obtiene de $P(T^+ \geq 9) = P(T^+ \leq 1) = 0.125$, para pruebas de cola derecha e izquierda,

respectivamente.

La media y varianza de T^+ son las siguientes:

$$\begin{aligned} \mathrm{E}(T^+) &= \frac{n(n+1)}{4} \\ \mathrm{var}(T^+) &= \frac{n(n+1)(2n+1)}{24} \end{aligned} \tag{7.12}$$

Por la generalización del *teorema Central del Límite*[2], la distribución asintótica de T^+ es normal con media *cero* y varianza *uno*, es decir,

$$\frac{T^+ - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{\mathrm{D}} Z \sim N(0,1) \tag{7.13}$$

donde $\xrightarrow{\mathrm{D}}$ significa que para muestras de tamaño suficientemente grande, la distribución de la estadística del lado izquierdo de (7.13) tiene aproximadamente la distribución normal estándar.

Si se rechaza $H_0 : \theta_D = 0$ frente a $H_1 : \theta_D > 0$ cuando $T^+ \geq k$, entonces el punto crítico k para la prueba de tamaño α se puede aproximar mediante

$$k \approx \frac{n(n+1)}{4} + 0.5 + Z_\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}}, \tag{7.14}$$

con 0.5 el valor de corrección por continuidad; Z_α es el percentil α de la distribución normal estándar.

La estadística de Wilcoxon se propone también para *datos pareados* con los cuales se hace inferencia sobre la mediana de la diferencia poblacional. Dada una muestra aleatoria de n pares $(X_1, Y_1) \cdots (X_n, Y_n)$, sus respectivas diferencias son $D_1 = X_1 - Y_1, \cdots, D_n = X_n - Y_n$; se asumen independientes de una población continua y simétrica con mediana θ_D para verificar la hipótesis

$$H_0 : \ \theta_D = \theta_0 \tag{7.15}$$

se asignan rangos al valor absoluto de las n diferencias $D_i = X_i - Y_i - \theta_0$, manteniendo el signo de la diferencia original. Los procedimientos para

[2]Combinación lineal de variables aleatorias *iid* como $B(1, 1/2)$.

estimación y pruebas de hipótesis arriba señalados se aplican acordemente, teniendo presente que el parámetro θ_D es ahora la mediana de las diferencias.

A manera de ejemplo, suponga que se registra el número de pacientes que por servicio de urgencias se atiende mensualmente en un hospital[3]. De sus registros se toma una muestra aleatoria de 15 meses diferentes, y se quiere verificar la hipótesis que la mediana del número de pacientes atendidos por urgencias es mayor de 80. Los datos para los 15 meses examinados son:

$$X_i : 62,\ 89,\ 82,\ 75,\ 76,\ 81,\ 83,\ 87,\ 100,\ 54,\ 88,\ 102,\ 75,\ 79,\ 64$$

La hipótesis a contrastar es

$$H_0 :\ \theta \leq 80 \text{ frente a } H_0 :\ \theta > 80$$

La diferencias $D_i = X_i - \theta_0$, los rangos de los $|D_i|$, los valores $R_i s(D_i)$ y el valor de la estadística T^+ (expresión (7.11)) son presentados en el siguiente recuadro.

D_i :	-18	9	2	-5	-4	1	3	7	20	-26	8	22	-5	-1	-16	Suma
$\|D_i\|$:	18	9	2	5	4	1	3	7	20	26	8	22	5	1	16	
R_i :	12	10	3	6.5	5	1.5	4	8	13	15	9	14	6.5	1.5	11	120
$R_i s(D_i)$:	0	10	3	0	0	1.5	4	8	13	0	9	14	0	0	0	62.5

Note que $T^+ = 62.5$ no es un valor extremo en relación con el máximo posible $T^+ + T^- = 120$; por tanto, los datos no provocan el rechazo de la hipótesis nula[4]. Se puede afirmar que el número medio de pacientes que son atendidos por servicio de urgencias en el hospital estudiado es a lo más 80.

[3] Se trata de una variable aleatoria con distribución de Poisson.

[4] Textos como el de Gil & Zárate (1984) disponen de tablas asociadas a la distribución de T^+.

7.3 Pruebas de localización en dos muestras

Aunque en la sección anterior se trabajó con datos de dos muestras, cada dato de una muestra está ligado a otro de la segunda muestra por alguna característica común (bloque). El esquema de muestreo que puede considerarse es de muestras dependientes o un muestreo de pares de una población bivariada. En esta parte se trabaja sobre muestras aleatorias extraídas de dos poblaciones independientes. Hay entonces dos niveles de independencia: por una parte, la independencia *dentro* de cada una de las observaciones de la muestra, y por la otra, la independencia *entre* las observaciones de una muestra a la otra.

Se presentarán algunos métodos basados en rangos para verificar y estimar diferencias en los parámetros de localización cuando las dos poblaciones tienen la misma forma.

Por comodidad, se notan las poblaciones por F_X y F_Y, respectivamente. Sean $X_1, X_2, \cdots, X_{n_1}$ y $Y_1, Y_2, \cdots, Y_{n_2}$ muestras aleatorias independientes y extraídas de F_X y F_Y, respectivamente.

La hipótesis de interés es

$$H_0 : \ \theta_X = \theta_Y \ \ \text{o} \ \ H_0 : \ \Delta = \theta_X - \theta_Y = 0 \tag{7.16}$$

7.3.1 Prueba de Mann-Whitney-Wilcoxon

Para verificar la hipótesis $H_0 : \ \Delta = 0$ frente a $H_1 : \ \Delta > 0$, se propone el siguiente procedimiento: mezclar las dos conjuntos de observaciones X y Y y ordenarlos en forma creciente, computar la estadística W que corresponde a la suma de los rangos de las Y[5], rechazar H_0 si W es grande, lo cual indica que los Y están desplazados a la derecha de los X. El problema entonces es encontrar la distribución de W bajo la hipótesis nula $H_0 : \ \Delta = 0$, para determinar la respectiva región crítica.

Sean $R_1, R_2, \cdots, R_{n_2}$ los rangos de las observaciones $Y_1, Y_2, \cdots, Y_{n_2}$. La variable aleatoria R_i tiene una distribución uniforme discreta, luego es inmediato que, bajo H_0, $\mathcal{E}(R_i) = (N+1)/2$, $var(R_i) = (N^2-1)/12$ y

[5]Un resultado equivalente se obtendría si se hace respecto a las X.

$cov(R_i, R_j) = -(N+1)/12$, para $i \neq j$. Como $W = \sum_{i=1}^{n_2} R_i$, entonces

$$\begin{aligned} \mathrm{E}(W) &= \frac{n_2(N+1)}{2} \text{ y} \\ \mathrm{var}(W) &= \frac{n_1 n_2 (N+1)}{12} \end{aligned} \tag{7.17}$$

Recuerde que R_i es el número de observaciones combinadas menores que o iguales a Y_i, es decir,

$$R_i = \#(X_j < Y_i) + \#(Y_k \leq Y_i), \quad j = 1, 2, \cdots, n_1 \quad k = 1, 2, \cdots, n_2$$

La estadística W se puede escribir en la forma

$$\begin{aligned} W &= \sum_{i=1}^{n_2} R_i = \#(Y_i - X_j > 0) + n_2(n_2+1)/2 \\ &= U + n_2(n_2+1)/2 \end{aligned} \tag{7.18}$$

donde

$$U = \#(Y_i - X_j) > 0) = \sum_{i=1}^{n_2} R_i - n_2(n_2+1)/2, \tag{7.18a}$$

es la estadística de *Mann-Whitney*, que cuenta el número de observaciones de X que son "superadas" por las Y. Note que

$$\begin{aligned} \mathrm{E}(U) &= \frac{n_1 n_2}{2} \\ \mathrm{var}(U) &= \frac{n_1 n_2 (N+1)}{12} \end{aligned} \tag{7.19}$$

Con la media y la varianza para W y U mostradas en (7.17) y (7.19), tan solo se requiere su distribución para construir pruebas de hipótesis e intervalos de confianza.

Considere $n_1 = 3$ y $n_2 = 2$ y suponga que la hipótesis nula es cierta. Hay $\binom{5}{2} = \binom{5}{3} = 10$ arreglos igualmente probables, listados en la tabla 7.2.

Se nota que U es de libre distribución, pues no se requiere conocer la distribución particular F para el cálculo de las probabilidades de U. Además, se puede observar que la distribución de U es simétrica respecto a $n_1 n_2/2$; en este caso, en torno a 3 (valor enmarcado en la tabla 7.2).

Tabla 7.2: Distribución de U con $n_1 = 3$ y $n_2 = 2$.

	Rangos							
	1	2	3	4	5	U	W	$f_U(u) = P(U = u)$
Arreglos								
	y	y	x	x	x	0	3	1/10
	y	x	y	x	x	1	4	1/10
	y	x	x	y	x	2	5	
	x	y	y	x	x	2	5	2/10
	y	x	x	x	y	3	6	
	x	y	x	y	x	3	6	2/10
	x	y	x	y	x	4	7	
	x	x	y	y	x	4	7	2/10
	x	x	y	x	y	5	8	1/10
	x	x	x	y	y	6	8	1/10

Para otros valores de n_1 y de n_2 se han elaborado tablas que contienen algunos cuantiles asociados a la distribución de U para valores de α específicos, generalmente 0.10, 0.05, 0.025 0.01 y 0.005 (Conover 1990, 384–388).

Los tres tipos de hipótesis alternativas y las respectivas regiones de rechazo se muestran en la tabla 7.3.1, donde U^* es "el número de veces que un Y precede a un X". Para encontrar el valor de k_α, para cualquier par de tamaños muestrales n_1 y n_2, se pueden enumerar los casos empezando con $u = 0$ y trabajar hasta que al menos $\alpha\binom{n_1+n_2}{n_1}$ casos sean contados.

La decisión de rechazar o no H_0 en un nivel de significación α depende de la magnitud de U y de la hipótesis (a), (b) o (c) que se verifique. A continuación se indican los respectivos criterios:

Tabla 7.3: Hipótesis alternativas y regiones de rechazo, prueba de Mann-Whitney.

	Alternativa	Región de rechazo
(a):	$\theta_X < \theta_Y$	$U \leq k_\alpha$
(b):	$\theta_X > \theta_Y$	$U^* \leq k_\alpha$
(c):	$\theta_X \neq \theta_Y$	$U \leq k_\alpha/2$ o $U^* \leq k_\alpha/2$

a) Para la alternativa $\theta_X < \theta_Y$, valores suficientemente grandes de U provocan el rechazo de $H_0 : \ \theta_X \geq \theta_Y$.

b) Para la alternativa $\theta_X > \theta_Y$, valores suficientemente pequeños de U provocan el rechazo de $H_0 : \ \theta_X \leq \theta_Y$.

c) Para la alternativa $\theta_X \neq \theta_Y$, valores suficientemente grandes o suficientemente pequeños de U provocan el rechazo de $H_0 : \ \theta_X = \theta_Y$.

Cuando n_1 y n_2 son grandes, las tablas para los valores críticos de la estadística U se vuelven incómodas o engorrosas de calcular; esto se obvia con la distribución asintótica (muestras de tamaño grande) de la estadística U. Como U es la suma de variables aleatorias idénticamente distribuidas (aunque dependientes), una generalización del Teorema Central del Límite permite afirmar que, bajo la hipótesis nula, la variable U estandarizada se aproxima a la normal estándar cuando $n_1, n_2 \to \infty$, siempre que $n_1/n_2 \longrightarrow \lambda$ (permanezca constante). En resumen, de acuerdo con (7.19),

$$\frac{U - \mathcal{E}(U)}{\sqrt{var(U)}} = \frac{U - n_1 n_2/2}{\sqrt{n_1 n_2 (N+1)/12}} \xrightarrow{D} Z \sim N(0,1) \tag{7.20}$$

donde $\xrightarrow{D}$ significa que, para muestras de tamaño suficientemente grande, la distribución de la estadística del lado izquierdo de la igualdad (7.20) tiene aproximadamente la distribución normal estándar.

Los datos que se muestran a continuación corresponden al tiempo (en minutos) necesario para la coagulación de la sangre, registrado en dos grupos de pacientes con el mismo perfil. Al primer grupo se le aplicó el medicamento a una dosis determinada; al otro, una dosis igual al doble de la del primero.

Grupo 1 (X):	6.31	6.33	9.90	6.28	6.50	6.73	
Grupo 2 (Y):	6.46	4.62	4.30	4.35	6.50	5.12	1.92

En este caso $n_1 = 6$ y $n_2 = 7$. Las diferencias $Y_i - X_j$ y sus rangos (entre paréntesis) están contenidos en la tabla siguiente:

Y_i	$Y_i-6.31$	$Y_i-6.33$	$Y_i-9.90$
6.46	0.15 (38)	0.13 (37)	−3.44 (11)
4.62	−1.69 (26)	−1.71 (25)	−5.28 (4)
4.30	−2.01 (19)	−2.03 (18)	−5.60 (2)
4.35	−1.96 (22)	−1.98 (20.5)	−5.55 (3)
6.50	0.19 (41)	0.17 (39)	−3.40 (12)
5.12	−1.19 (31)	−1.21 (30)	−4.78 (6)
1.92	−4.39 (9)	−4.41 (8)	−7.98 (1)
Y_i	$Y_i-6.28$	$Y_i-6.50$	$Y_i-6.73$
6.46	0.18 (39)	−0.04 (35)	−0.27 (33)
4.62	−1.66 (27)	−1.88 (24)	−2.11 (17)
4.30	−1.98 (20.5)	−2.20 (15)	−2.43 (13)
4.35	−1.93 (23)	−2.15 (16)	−2.38 (14)
6.50	0.22 (42)	0.00 (36)	−0.23 (34)
5.12	−1.16 (32)	−1.38 (29)	−1.61 (28)
1.92	−4.36 (10)	−4.58 (7)	−4.81 (5)

Las muestras ordenadas con sus respectivos rangos están en el siguiente cuadro[6]:

Y_i	X_j	$R(Y_i)$	$R(X_j)$
1.92		1	
4.30		2	
4.35		3	
4.62		4	
5.12		5	
	6.28		6
	6.31		7
	6.33		8
6.46		9	
	6.50		10.5
6.50		10.5	
	6.73		12
	9.90		13
		$\sum_{i=1}^{7} R(Y_i) = 34.5$	$\sum_{i=1}^{6} R(X_j) = 56.5$

[6]La función `wilcox.test()` de `R` resta la menor de las observaciones antes de calcular los rangos

De la ecuación (7.18),

$$\begin{aligned} U &= W - n_2(n_2+1)/2 = \sum_{i=1}^{n_2} R_i - n_2(n_2+1)/2 \\ &= 34.5 - \frac{7\times 8}{2} \\ &= 6.5 \end{aligned}$$

Se puede concluir, de acuerdo con la tabla (7.3.1), que hay diferencia significativa entre los parámetros de localización para las dos poblaciones; es decir, que la dosis influye de manera diferente en el tiempo de coagulación de la sangre en este tipo de personas.

La tabla 7.4 contiene los datos de un estudio sobre la relación entre el consumo de cloruro de sodio y la hipertensión. Dos grupos de sujetos, 12 normales y 10 hipertensos, fueron aislados durante una semana y se compararon en relación con el promedio de ingesta diaria de Na^+.

Tabla 7.4: Consumo de cloruro de sodio.

Grupo normal		Grupo hipertenso	
Sujeto	Na^+	Sujeto	Na^+
1	10.2	1	92.8
2	2.2	2	54.8
3	0.0	3	51.6
4	2.6	4	61.7
5	0.0	5	250.8
6	43.1	6	84.5
7	45.8	7	34.7
8	63.6	8	62.2
9	1.8	9	11.0
10	0.0	10	39.1
11	3.7		
12	0.0		

Mediante el procedimiento `NPAR1WAY` del SAS se computa la estadística de Wilcoxon. Una vez ingresados los datos, el procedimiento SAS es el siguiente:

```
PROC NPAR1WAY wilcoxon;
CLASS grupo;
VAR consumo;
RUN;
```

El siguiente cuadro contiene la salida del procedimiento `NPAR1WAY` para los datos de la tabla 7.4.

```
Wilcoxon Scores (Rank Sums) for Variable CONSUMO
Classified by Variable GRUPO

               Sum of  Expected  Std  Dev    Mean
     GRUPO  N  Scores  Under H0  Under H0    Score

    Normal 12     91.0    138.0  15.1228734  7.5833333
Hipertenso 10    162.0    115.0  15.1228734 16.2000000

Average Scores Were Used for Ties
Wilcoxon 2-Sample Test (Normal Approximation)
(with Continuity Correction of .5)

S=162.000    Z = 3.07481   Prob > |Z| = 0.0021
```

7.4 Pruebas de localización en diseños completamente al azar

Se trata ahora de extender el problema estudiado en las dos primeras partes de este capítulo, es decir, considerar los problemas de localización para una y dos muestras en k muestras independientes, es decir, *análisis de varianza*.

Se trabaja en una estructura de diseño a *una vía de clasificación* para este tipo de datos. Mediante las k muestras se quiere verificar la hipótesis de que los datos provienen de una misma población (la media no difiere significativamente).

Una situación experimental, para este caso, es aquella en que k muestras aleatorias han sido obtenidas de k poblaciones, posiblemente diferentes. Se quiere verificar la hipótesis que todas las poblaciones son idénticas

frente a la alternativa de que algunas poblaciones tienden a poseer valores más grandes (o pequeños) que otras.

Se presentará con especial atención la prueba de *Kruskal-Wallis*, junto con una prueba de comparaciones múltiples, para ayudar a identificar las poblaciones que provocan la diferencia.

7.4.1 Prueba de Kruskal-Wallis

Kruskal y Wallis (1952) presentan una prueba para arreglos de una vía de clasificación. El diseño de muestreo consiste en k muestras

$$X_{11}, X_{21}, \cdots, X_{n_1 1}, \quad X_{12}, X_{22}, \cdots, X_{n_2 2}, \cdots, X_{1k}, X_{2k}, \quad \cdots, X_{n_k k}$$

de poblaciones $F(x, \theta_1), F(x, \theta_2), \cdots, F(x, \theta_k)$, respectivamente. Una disposición de los datos más cómoda de leer es la siguiente:

Muestra						
1	2	$\cdots$	j	$\cdots$	k	
X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1k}	
X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2k}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$X_{n_1 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_j j}$	$\cdots$	$X_{n_k k}$	
$X_{.1}$	$X_{.2}$	$\cdots$	$X_{.j}$	$\cdots$	$X_{.k}$	$X_{..}$

donde X_{ij} es la i-ésima replicación en la j-ésima muestra (tratamiento), $X_{.j} = \sum_{i=1}^{n_j} X_{ij}$ es el total en la j-ésima muestra y $X_{..} = \sum_{i=1}^{n_j} \sum_{j=1}^{k} X_{ij}$ es el gran total.

En forma análoga, la media dentro de la j-ésima muestra es

$$\bar{X}_{.j} = \frac{X_{.j}}{n_j}$$

y la media general

$$\bar{X}_{..} = \frac{X_{..}}{N},$$

con $N = \sum_{j=1}^{k} n_j$, el total de observaciones.

Se quiere construir una prueba para la hipótesis:

$$H_0 : \ \theta_1 = \theta_2 = \cdots = \theta_k \text{ frente a } H_1 : \ \theta_1, \theta_2, \cdots, \theta_k \text{ no todas iguales} \tag{7.21}$$

Las técnicas no paramétricas desarrolladas para el problema de k muestras no requieren otro supuesto más que el de continuidad. La estrategia básica de la prueba de Kruskal-Wallis es asignar rangos a las N observaciones y comparar la suma de los rangos por muestra (columna o tratamiento). Sea R_{ij} el rango de X_{ij}. La tabla equivalente a la anterior para rangos es:

Muestra						
1	2	$\cdots$	j	$\cdots$	k	
R_{11}	R_{12}	$\cdots$	R_{1j}	$\cdots$	R_{1k}	
R_{21}	R_{22}	$\cdots$	R_{2j}	$\cdots$	R_{2k}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$R_{n_1 1}$	$R_{n_2 2}$	$\cdots$	$R_{n_j j}$	$\cdots$	$R_{n_k k}$	
$R_{.1}$	$R_{.2}$	$\cdots$	$R_{.j}$	$\cdots$	$R_{.k}$	$R_{..}$

donde:

$$R_{.j} = \sum_{i=1}^{n_j} R_{ij}, \text{ y } \bar{R}_{.j} = \frac{R_{.j}}{n_j} \tag{7.22}$$

La estadística de *Kruskal-Wallis* se expresa por

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{1}{n_j} \left[R_{.j} - \frac{n_j(N+1)}{2} \right]^2 \tag{7.23}$$

Bajo la hipótesis, que las muestras provienen de la misma población, H tiene una distribución asintótica *ji–cuadrado* con $(k-1)$ grados de libertad.

La siguiente expresión es equivalente algebraicamente a la contenida en (7.23)

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_{.j}^2}{n_j} - 3(N+1) \tag{7.24}$$

Se rechaza $H_0 : \ \theta_1 = \theta_2 = \cdots = \theta_k$ con un nivel de significancia aproximado de α cuando $H \geq \chi^2_\alpha(k-1)$, donde $\chi^2_\alpha(k-1)$ es el percentil $(1-\alpha)$ de la distribución ji–cuadrado con $k-1$ grados de libertad.

El supuesto inicial de población continua obvia, al menos teóricamente, el problema de empates. Por problemas muestrales se pueden presentar empates, esta situación se resuelve con el método del rango promedio explicado anteriormente. En caso de empates se sigue el mismo procedimiento de la prueba de Mann-Whitney. La siguiente es la estadística con corrección para empates

$$H^* = \frac{H}{\left[1 - \sum_{i=1}^{r} \frac{e_i(e_i^2-1)}{N(N^2-1)}\right]} \tag{7.25}$$

donde r es el número total de empates y e_i el número de observaciones empatadas en el i-ésimo empate.

En un estudio experimental se quieren comparar tres dietas, con un contenido de colesterol diferente, en cuanto a la tensión arterial sistólica (TAS) en personas de edad avanzada. Se consideraron 5 personas con la dieta tipo A, 4 con la tipo B y 4 con la C. A continuación se presentan los datos junto con sus respectivos rangos.

	Dieta		
	A	B	C
	172 (9.5)	175 (11)	160 (4.5)
	169 (8.0)	164 (7.0)	160 (4.5)
	180 (13.0)	150 (2.5)	150 (2.5)
	172 (9.5)	161 (6.0)	148 (1.0)
	178 (12.0)		
$R_{.j}$:	$R_{.1} = 52.0$	$R_{.2} = 26.5$	$R_{.3} = 12.5$

En este caso, $R_{..} = 52 + 26.5 + 12.5 = 91$, $N = 5 + 4 + 4 = 13$ y H toma el valor

$$\begin{aligned} H &= \frac{12}{(13)(14)}\left[\frac{52.0^2}{5} + \frac{26.5^2}{4} + \frac{12.5^2}{4}\right] - 3(14) \\ &= 7.808 \end{aligned}$$

De la tabla (A.2), para $n_1 = 5$, $n_2 = n_3 = 4$, $P(H \geq 7.7604) = 0.009$, como 7.808 es un valor más extremo, su valor p será menor que 0.009, es decir, $\alpha^* = P(H \geq 7.808) < 0.0094$.

Con el valor $\chi^2_{0.05} = 5.9915$ se rechaza también la hipótesis de que los TAS promedio son iguales para las tres dietas.

El cálculo de H^*, debido al problema de empates, es el siguiente. En este caso hay tres empates (en 150, 160 y 172), $r = 2$. Los tres tienen de a dos empates, de donde $e_1 = e_2 = e_3 = 2$, de acuerdo con (7.25)

$$H^* = \frac{7.808}{\left[1 - \frac{2(4-1)+2(4-1)+2(4-1)}{(13)(168)}\right]} = 7.866$$

H^* no cambia sustancialmente la decisión anterior.

Para efectos de cálculo, con el procedimiento `NPAR1WAY` del paquete SAS se puede desarrollar el análisis de varianza, que lleva al cálculo de la estadística de Kruskal-Wallis. Otra alternativa consiste en asignar rangos a las observaciones mediante el procedimiento RANK del SAS, y luego desarrollar un análisis de varianza corriente mediante el procedimiento ANOVA (o el GLM) del mismo SAS. Al finalizar este capítulo se presenta la sintaxis para el caso de diseños en bloques. En R se cuenta con la función `kruskal.test()`; en la sección 7.6 se presentan detalles de su uso.

7.5 Pruebas de localización para diseños en bloques aleatorizados completos

En la sección (7.4) se compararon k muestras para detectar diferencias de localización entre las poblaciones muestreadas. Se asignaron aleatoriamente los k tratamientos a las N unidades experimentales. A veces ocurre que la diferencia entre tratamientos resulta ocultada por una variabilidad relativa grande de las unidades experimentales en la muestra. Una estrategia para "curarse" o "prevenirse" de este problema es dividir las unidades experimentales en subgrupos o *bloques* homogéneos. La homogeneidad en cada bloque se hace respecto a la variable o factor que se considere puede enmascarar la verdadera diferencia entre las poblaciones (tratamientos). Una vez que se ha decidido la variable "extraña" a controlar, se conforman los bloques respecto a varios niveles de esta variable. Finalmente, en cada uno de estos bloques se asignan aleatoriamente los tratamientos. De aquí el calificativo de *bloques aleatorizados completos*.

Las medidas repetidas sobre un mismo sujeto son un caso especial de bloqueo; cada individuo hace de bloque, al cual se le aplican aleatoriamente varios tratamientos como medicamentos, terapias, conceptos, métodos quirúrgicos, etc. (Díaz 2007).

7.5.1 Prueba de Friedman

Considere la muestra aleatoria X_{ij}, con $i = 1, \cdots, b$ y $j = 1, \cdots, k$ de una población $F_i(x, \theta_j)$. Es decir, una muestra aleatoria de tamaño k sobre cada uno de los b bloques.

Se desea verificar $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$, frente a H_1 : "no todos los θ_i son iguales". El procedimiento consiste en asignar rangos R_{ij} a cada una de las observaciones X_{ij} entre las observaciones $X_{i1}, X_{i2}, \cdots, X_{ik}$ de cada bloque $i = 1, 2, \cdots, b$. Se escribe $R_{.j} = \sum_{i=1}^{b}$ al total de rangos del j-ésimo tratamiento.

Una presentación tabular de los rangos correspondientes a cada una de las b muestras es la siguiente:

	Tratamiento						
Bloque	1	2	$\cdots$	j	$\cdots$	k	
1	R_{11}	R_{12}	$\cdots$	R_{1j}	$\cdots$	R_{1k}	
2	R_{21}	R_{22}	$\cdots$	R_{2j}	$\cdots$	R_{2k}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
i	R_{i1}	R_{i2}	$\cdots$	R_{ij}	$\cdots$	R_{ik}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
b	R_{b1}	R_{b2}	$\cdots$	R_{bj}	$\cdots$	R_{bk}	
	$R_{.1}$	$R_{.2}$	$\cdots$	$R_{.j}$	$\cdots$	$R_{.k}$	$R_{..}$

Bajo H_0, los rangos $R_{i1}, R_{i2}, \cdots, R_{ik}$ y $R_{.j}$ se distribuyen conforme se presentó en la sección (7.2), es decir,

$$\begin{aligned} \mathrm{E}(R_{ij}) &= \frac{k+1}{2}, & \mathrm{var}(R_{ij}) &= \frac{(k^2-1)}{12} \\ \mathrm{E}(R_{.j}) &= \frac{b(k+1)}{2}, & \mathrm{var}(R_{.j}) &= \frac{b(k^2-1)}{12} \\ cov(R_{.j}, R_{.j'}) &= -\frac{b(k+1)}{2} \end{aligned} \tag{7.26}$$

La expresión para la estadística de Friedman es la siguiente:

$$K = \sum_{j=1}^{k} c_{jN}^2 \left\{ \frac{R_{.j} - \mathcal{E}(R_{.j})}{\sqrt{var(R_{.j})}} \right\}^2 \tag{7.27}$$

donde las constantes se escogen en forma tal que K tenga una distribución asintótica $\chi^2_{(k-1)}$. En Hettmansperger (1984, 197), se muestra la forma apropiada de las constantes. Entonces la estadística de Friedman toma la forma

$$\begin{aligned} K^* &= \sum_{j=1}^{k} \Big(1 - \frac{1}{k}\Big) \left\{ \frac{R_{.j} - b(k+1)/2}{\sqrt{b(k^2-1)/12}} \right\}^2 \\ &= \frac{12}{bk(k+1)} \sum_{j=1}^{k} \Big[R_{.j} - b(k+1)/2 \Big]^2 \\ &= \left[\frac{12}{bk(k+1)} \sum_{j=1}^{k} R_{.j}^2 \right] - 3b(k+1) \end{aligned} \tag{7.28}$$

Se rechaza la hipótesis nula $H_0 : \ \theta_1 = \theta_2 = \cdots = \theta_k$ en un nivel α, si $K^* \geq \chi^2_{\alpha,(k-1)}$, donde $\chi^2_{\alpha,(k-1)}$ es el percentil $1-\alpha$ de una distribución ji–cuadrado de $k-1$ grados de libertad.

Suponga que se desarrolló un experimento en bloques aleatorizados completos para verificar cuatro medicamentos en niños recién nacidos. Cada bloque se conforma por niños que están dentro de un mismo rango de peso. Las observaciones corresponden a la ganancia en hierro en la sangre al cabo de una semana (en mg/cc).

	Peso (Bloque)			
Medicamento	1	2	3	Total ($R_{.j}$)
A	10.2 (4)*	11.3 (4)	10.7 (4)	32.2 (12)
B	9.7 (3)	10.8 (3)	9.9 (3)	30.4 (9)
C	7.7 (1)	8.4 (1)	8.8 (2)	24.9 (4)
D	8.1 (2)	9.0 (2)	8.7 (1)	25.8 (5)
Total	35.7	39.5	38.1	113.3

$*$ rango de la observación respecto a las otras del mismo bloque.

El valor de la estadística K^* de Friedman es

$$\begin{aligned} K^* &= \left[\frac{12}{bk(k+1)}\sum_{j=1}^{k} R_{.j}^2\right] - 3b(k+1) \\ &= \left[\frac{12}{(3)(4)(4+1)}\{12^2+9^2+4^2+5^2\}\right] - 3(3)(4+1) \\ &= 8.2 \end{aligned}$$

Como este último valor es mayor que $\chi_{0.05}(3) = 7.81$, se puede afirmar, con un nivel de significancia del 5%, que hay diferencias respecto en el aumento en hierro promedio para los cuatro medicamentos.

En el siguiente cuadro se muestra la sintaxis para desarrollar la prueba de Friedman, con los procedimientos RANK y ANOVA del paquete SAS. En la sección 7.6 se explica cómo hacerlo en R.

```
TITLE  'DISEÑO EN   BLOQUES  ALEATORIZADOS';
 DATA A1;
INPUT  BLOQUE  TRATAM$ RESPUES;
CARDS;
 1  A  10.2   1  B  9.7   1  C  7.7   1  D  8.1
 2  A  11.3   2  B 10.8   2  C  8.4   2  D  9.0
 3  A  10.7   3  B  9.9   3  C  8.8   3  D  8.7
 ;
PROC  RANK; /* procedimiento para asignar rangos */
BY  BLOQUE;
VAR  RESPUES;
RANKS  R_RESPU;
RUN;
PROC  ANOVA;  */ procedimiento para el análisis de
                 varianza */
CLASS BLOQUE  TRATAM;
MODEL  RESPUES  =  BLOQUE  TRATAM;
    TITLE2  'PRUEBA  DE  FRIEDMAN';
    RUN;
```

7.6 Procesamiento de datos con R

En esta sección se explica cómo realizar las diferentes pruebas tratadas en este capítulo mediante el programa R. Se estudia la funcion `wilcox.test()` para realizar la prueba de rangos signados de Wilcoxon y la prueba de Mann–Whitney–Wilcoxon; para realizar la prueba de Kruskal–Wallis, la función `kruskal.test()` y, por último, la función `friedman.test()` para llevar a cabo la prueba de Friedman.

Prueba del signo

Prueba del signo, una sola muestra

No existe, en el paquete básico de R, una función para realizar la prueba del signo; sin embargo su implementación en código de R es relativamente sencilla. A continuación se ilustra como hacerlo con los ejemplos tratados en este capítulo, concretamente en la sección 7.2.1.

```
# valor especificado por la hipótesis nula
mediana<-5
# se introducen los datos
x<-c(7.1,1.2,1.9,2.5,6.1,0.2,1.2,0.2,1.1,
     1.6,0.3,10.4,3.1,13.3,8.1,5.3)
# numero de observaciones
n<-length(x)
# si la observación es mayor que la mediana regresa 1
# en otro caso 0 luego se suman
S<-sum(ifelse(x-mediana>0,1,0) )
# p--valor
pbinom(S,n,1/2)
```

Prueba del signo, muestras pareadas

A continuación se ilustra cómo realizar la prueba del signo con muestras pareadas usando los datos del ejemplo de la sección 7.2.1.

```
# peso antes
pa<-c(80,82,75,90,98,87,100,107,103)
```

```
# peso después
pd<-c(79,83,75,81,95,86,101,105,100)
# diferencias
dif<-pa-pd
# se eliminan los empates
dif<-dif[dif!=0]
# numero de observaciones (sin empates)
n<-length(dif)
# número de diferencia
S<-sum(ifelse(dif>0,1,0))
# p--valor
pbinom(n-S,n,1/2)
```

Prueba de rango signado de Wilcoxon

La prueba de rango signado de Wilcoxon, se realiza con la función `wilcox.test()`, la cual se explica en esta sección mediante un ejemplo. Antes de estudiar esta función, se explica cómo obtener rangos.

En R, para asignar rangos, se usa la función `rank()`. Veamos un ejemplo:

```
x<-c(4.0,8.6,3.2,2.0,1.9,0.5,5.8)
R<-rank(x)
```

Para ilustrar el uso de la función `wilcox.test()` se utilizan los datos del ejemplo que se encuentra al final de la sección 7.2.3. A continuación se muestra el código y su salida:

```
# introducción de datos
x<-c(62,89,82,75,76,81,83,87,100,54,88,102,54,75,79,64)
wilcox.test(x,alternative ="greater",mu=80,exact = FALSE)

Wilcoxon signed rank test with continuity correction

data:  x
V = 62.5, p-value = 0.6219
alternative hypothesis: true location is greater than 80
```

En la salida anterior, el valor `V = 62.5` corresponde a la estadística T^+ (ecuación (7.11)). La opción `alternative` indica la hipótesis alterna-

tiva, (para este ejemplo: $H_1 : \theta > 80$); los otros dos valores posibles son `less`, para $H_1 : \theta < 80$, y `two.sided`, para $H_1 : \theta \neq 80$. Esta última es la opción por defecto. La opción `mu=80` indica el valor de la mediana bajo la hipótesis nula; por defecto su valor es cero. La opción `exact = FALSE` permite que el $p-$valor se obtenga mediante la aproximación por la distribución normal; si se usa `exact = TRUE`, este valor se calcula usando la distribución exacta de la estadística T^+. En este caso la dificultad está cuando se presentan empates, caso en que no es posible obtener la distribución exacta. Observe en la salida anterior que por defecto se está usando la corrección por continuidad (7.14); si no se desea la corrección, debe usarse la opción `correct=FALSE`, como se muestra a continuación:

```
x<-c(62,89,82,75,76,81,83,87,100,54,88,102,54,75,79,64)
wilcox.test(x,alternative ="greater",mu=80,exact = FALSE,
            correct=FALSE)
```

Otra opción importante de esta función es `conf.int`, la cual, si se fija en `TRUE`, entrega un intervalo de confianza al 95% para la mediana. El nivel de confianza se puede cambiar mediante la opción `conf.level`.

Prueba de Mann–Whitney–Wilcoxon

Se ilustra la realización de esta prueba mediante el ejemplo que se encuentra al final de la sección 7.3.1, con los datos de la tabla 7.4.

```
norm<-c(10.2,2.2,0,2.6,0,43.1,45.8,63.6,1.8,0,3.7,0)
hiper<-c(92.8,54.8,51.6,61.7,250.8,84.5,34.7,62.2,
         11.0,39.1)
wilcox.test(norm,hiper,correct=TRUE,exact=FALSE)

        Wilcoxon rank sum test with continuity correction
data:  norm and hiper
W = 13, p-value = 0.002106
alternative hypothesis: true location shift is
not equal to 0
```

Las opciones discutidas en la sección anterior también son válidas en este caso.

Prueba de Kruskal–Wallis

Con el siguiente código se realiza la prueba de Kruskal–Wallis, tratada en la sección 7.4.1:

```
# se introducen los datos
tas<-c(172,169,180,172,178,175,164,150,161,160,
       160,150,148)
# definición del factor que identifica cada tratamiento
dieta<-factor( rep(1:3,c(5,4,4)), labels=c("A","B","C"))
# prueba
kruskal.test(tas~dieta)

# salida de la función
        Kruskal-Wallis rank sum test

data:  tas by dieta
Kruskal-Wallis chi-squared = 7.8731,
df = 2, p-value = 0.01952
```

Otra forma de entregar los datos a la función `kruskal.test()` es crear vector para cada dieta y organizarlos luego en una lista, de la siguiente forma:

```
# dieta A
A<-c(172,169,180,172,178)
# dieta B
B<-c(175,164,150,161)
# dieta C
C<-c(160,160,150,148)
kruskal.test(list(A,B,C))
```

Prueba de Friedman

El entorno y lenguaje `R` cuenta con la función `friedman.test()` para realizar la prueba de Friedman, cuyo uso se ilustra a continuación, mediante los datos del ejemplo al final de la sección 7.5.1.

```
y<-c(10.2,11.3,10.7, 9.7,10.8,9.9, 7.7,8.4,8.8,
         8.1,9.0,8.7)
```

```
medic<-rep(c("A","B","C","D"),rep(3,4))
bloque<-rep(1:3,4)
datos<-data.frame(y,medic,bloque,R=rank(y))
friedman.test(y~medic|bloque,data=datos)

# salida de la función
        Friedman rank sum test

data:  y and medic and bloque
Friedman chi-squared = 8.2, df = 3, p-value = 0.04205
```

Otra forma de entregar los datos es mediante una matriz cuyas filas contienen la información de los tratamientos y las columnas corresponden a los bloques:

```
friedman.test(t(matrix(y,nrow=4,byrow=TRUE)))
```

7.7 Ejercicios

1. Los tiempos de sobrevivencia (en años) de 12 personas que se han sometido a un trasplante de corazón son 3.1, 0.9, 2.8, 4.3, 0.6, 1.4, 5.8, 9.9, 6.3, 10.4, 0, 11.5. Probar si los datos sugieren que la mediana es mayor de 5 años.

2. Con el objetivo de establecer la influencia de un fármaco en el tratamiento de las várices, una de las variables que se estudió fue la variación del peso del paciente después del tratamiento. Los resultados obtenidos sobre una muestra de 32 pacientes se observan en la tabla 7.5. ¿Hubo una alteración significativa en el peso de los pacientes tratados? Justifique.

3. En un estudio realizado para averiguar el tipo de población atendida en un centro hospitalario, se encontró que la mediana de edad de los enfermos era de 57 años. En un estudio similar de otro hospital se tomó una muestra de 15 personas cuyas edades se muestran a continuación: 26, 90, 44, 67, 12, 34, 67, 66, 24, 49, 45, 15, 58, 77, 57. Use la prueba de rango signado de Wilcoxon para decidir si la mediana de edad de los pacientes en ambos hospitales es igual.

Tabla 7.5: Datos sobre variación de pesos de pacientes tratados para várices.

Ind	Peso inicial	Peso final	Ind	Peso inicial	Peso final
1	73	69	17	85	85
2	99	93	18	94	94
3	75	78	19	89	88
4	84	85	20	57	54
5	102	99	21	59	56
6	84	80	22	67	65
7	65	67	23	96	95
8	70	72	24	97	99
9	78	73	25	73	75
10	75	71	26	58	56
11	78	78	27	57	57
12	82	80	28	63	62
13	64	61	29	81	80
14	72	74	30	84	85
15	71	76	31	80	78
16	64	63	32	67	68

4. Se llevó a cabo un estudio para determinar el grado en el cual el alcohol entorpece la habilidad de pensamiento para llevar a cabo determinada tarea. Se seleccionaron al azar 10 personas de distintas características. Cada persona llevó a cabo la tarea sin nada de alcohol en su organismo. Entonces, la tarea volvió a llevarse a cabo después de que cada persona había consumido una cantidad de alcohol para tener en su organismo un contenido de 0.1%. Los tiempos, antes y después (en minutos), se muestran en la tabla 7.6. ¿Puede concluirse que el tiempo promedio *antes* es menor que el tiempo promedio *después* por más de 10 minutos?

5. A finales de la década de los setenta se descubrió que la sustancia carcinogénica nitrosodimetilamina (NDMA) se formaba durante el secado de la malta verde, la cual se usaba para fabricar cerveza. A principios de los ochenta se desarrolló un nuevo proceso para el secado de la malta, el cual minimizaba la formación de NDMA. Se tomaron muestras aleatorias de una cerveza doméstica que se fabricó empleando ambos procesos de secado, y se midieron los

Tabla 7.6: Tiempos para desarrollar una tarea con o sin alcohol.

Participante	Antes	Después
1	28	39
2	22	45
3	55	67
4	45	61
5	32	46
6	35	58
7	40	51
8	25	34
9	37	48
10	20	30

niveles de NDMA en partes por billón. Se obtuvieron los datos que se muestran en la tabla 7.7. ¿Existe alguna razón para creer que ha disminuido la cantidad de NDMA en más de dos partes por billón con el empleo del nuevo proceso?

Tabla 7.7: Niveles de NDMA.

Proceso anterior	6	4	5	5	6	5	5	6	4	6	7	4
Proceso nuevo	2	1	2	2	1	0	3	2	1	0	1	3

6. Con el objetivo de verificar el contenido de alquitrán, se prueban muestras aleatorias de cuatro marcas de cigarros. Las cifras de la tabla 7.8 corresponden, en miligramos, al alquitrán encontrado en los 16 cigarros probados. Utilice la prueba de Kruskal-Wallis, en el nivel de significación de 0.05, para probar si existe una diferencia significativa en el contenido de alquitrán entre las cuatro marcas de cigarros.

7. Dieciséis individuos con sobrepeso participaron en un análisis para comparar cuatro dietas para reducción de peso. Los individuos formaron grupos de acuerdo con el peso inicial y a cada uno de los cuatros sujetos se le asignó, al azar, una de las cuatro dietas de reducción de peso. Al terminar el periodo se consideró la pérdida de peso en libras. Los datos, que se muestran en la tabla 7.9,

Tabla 7.8: Niveles de alquitrán.

Marcas			
A	B	C	D
14	16	16	17
10	18	15	20
11	14	14	19
13	15	12	21

Tabla 7.9: Reducción de peso en libras.

Peso	Régimen			
inicial (libras)	A	B	C	D
150 a 174	12	26	24	23
175 a 199	15	29	23	25
200 a 225	15	27	25	24
Más de 225	18	38	33	31

¿proporcionan suficiente evidencia para indicar una diferencia en los efectos de la dieta?

Capítulo 8

Métodos para datos de conteo

8.1 Introducción

Datos como el número de bacterias o virus en una colonia, el número de accidentes o la incidencia de una enfermedad, la mayoría de las veces tienen asociada una distribución de Poisson. El interés generalmente se dirige a la estimación de una tasa o una incidencia (bacterias por unidad de volumen de una solución o muertes mensuales por cáncer de personas expuestas a un cancerígeno) y a la determinación de la relación entre dicha tasa y un conjunto de variables explicativas.

En la sección 1.4.1 se muestra la distribución de Poisson, de la cual se retoman algunos aspectos importantes para este capítulo. El modelo de Poisson es usado cuando una variable X representa el número de ocurrencias de algún evento aleatorio en un intervalo de tiempo o espacio, o volumen de materia. La media y la varianza de una distribución de Poisson son iguales:

$$\begin{aligned} \mathrm{E}(X) &= \mu \\ \mathrm{var}(X) &= \sigma^2 = \mu \end{aligned} \tag{8.1}$$

Dada una muestra $X_1, X_2, \ldots, X_n$ de distribución de Poisson $P(\mu)$, que corresponden a conteos, la media muestral $\overline{X}$ es un estimador insesgado

para μ; su error estándar es dado por

$$\text{ee}(\overline{X}) = \sqrt{\frac{\overline{X}}{n}} \tag{8.2}$$

A manera de ilustración, se considera el caso de estimar la tasa de infección en una colonia de organismos. Resulta complicado o inoperante ensayar cada organismo de manera individual para establecer su condición de infectado o no; una alternativa consiste en dividir aleatoriamente la colonia de organismos en cajas o compartimientos y probar cada caja como una unidad (Le 1998, 205). Sea

N: el número de insectos en la colonia

n: el número de cajas en el experimento

m: el número de insectos por caja, $N = nm$ (por comodidad se asume que m es el mismo para cada caja)

La variable aleatoria X registra el número de cajas que muestran una prueba negativa (es decir, ninguno de los insectos está infectado).

Sea μ la tasa de infección poblacional. La probabilidad de que todos los m insectos en una caja sean negativos es dada por

$$\pi = (1 - \mu)^m,$$

que equivale a la probabilidad de tener una caja negativa.

Al considerar una caja negativa como un "éxito", la variable aleatoria X tiene una distribución de tipo binomial, es decir, $X \sim B(n, \pi)$ (sección 1.4.2).

En situaciones en las cuales la tasa de infección μ es muy pequeña, la distribución de Poisson se puede emplear para aproximar esta probabilidad, así:

$$P(X = 0) = \pi = \frac{e^{-\mu}\mu^0}{0!} = \exp[-\mu].$$

Se observa que una estimación de esta tasa involucra la estimación de μ.

Algunas pruebas sobre enfermedades como la sífilis y el VIH son aplicadas de esta manera.

8.2 Determinación de la naturaleza aleatoria de un evento

Una aplicación interesante de la distribución de Poisson tiene que ver con la ocurrencia aleatoria de un evento en un área geográfica determinada. En epidemiología, por ejemplo, es útil determinar si la presencia de una enfermedad en las personas que habitan un área definida está distribuida aleatoriamente o si esta tiene alguna tendencia no aleatoria de aglomeración espacial. Casos como estos se pueden encarar con una técnica que verifica el ajuste de un conjunto de datos, de una determinada área, a una distribución de Poisson.

Considere la siguiente situación: un epidemiólogo quiere verificar si una enfermedad está distribuida aleatoriamente o, por el contrario, se puede atribuir a algún patrón de contagio. Para esto divide adecuadamente el área geográfica de interés en zonas (cuadras, barrios, localidades, etc.) con aproximadamente la misma área y densidad poblacional, y selecciona aleatoriamente 100 de estas zonas. En cada una de las zonas cuenta el número de enfermos presentes. Suponga que este procedimiento arroja los siguientes datos:

$n_0 = 41$ zonas contienen 0 personas enfermas

$n_1 = 25$ zonas contienen 1 persona enferma

$n_2 = 20$ zonas contienen 2 personas enfermas

$n_3 = 14$ zonas contienen 3 personas enfermas

A continuación se esquematiza el procedimiento.

- El número total de enfermos se obtiene de multiplicar el número de zonas por el número de enfermos por zona y sumar: $41\times0+25\times1+20\times2+14\times3 = 0+25+40+42 = 107$. Así, una estimación de μ es $\overline{X} = 1.07$, es decir, 1.07 enfermos por zona. La probabilidad de que en una zona no hayan enfermos es $P(X = 0) = e^{[-1.07]} = 0.343$. De la expresión para la distribución de Poisson (ecuación (1.1) o tabla 1.5), se advierte que esta probabilidad es decreciente en tanto X sea mayor que 1; por esta razón, a la distribución de Poisson se le denomina distribución de los sucesos *"raros"* o *"aislados"*.

- Ahora se determinan las probabilidades de que cualquier zona elegida aleatoriamente contenga cero personas enfermas, una persona enferma, dos personas enfermas, y así sucesivamente. Estas serían las probabilidades que se tendrían si la distribución de los enfermos fuese aleatoria en el área de estudio, es decir, tipo Poisson. De esta forma:

$$
\begin{aligned}
P(\text{cero enfermos}) = P(X=0) &= e^{-1.07} = 0.343 \\
P(\text{un enfermo}) = P(X=1) &= e^{-1.07}(1.07) = 0.367 \\
P(\text{dos enfermos}) = P(X=2) &= \frac{e^{-1.07}(1.07)^2}{2!} = 0.196 \\
P(\text{tres enfermos}) = P(X=3) &= \frac{e^{-1.07}(1.07)^3}{3!} = 0.070
\end{aligned}
$$

- A continuación se determina el número de zonas que se puede esperar contengan cero, una, dos o tres personas enfermas si las personas se distribuyen aleatoriamente en el área. Para esto se multiplica cada valor de probabilidad encontrado en el paso anterior por el total de zonas, en este caso 100. Así,

$$
\begin{aligned}
100 \times 0.343 &= 34.3 = E_0 \\
100 \times 0.367 &= 36.7 = E_1 \\
100 \times 0.196 &= 19.6 = E_2 \\
100 \times 0.070 &= 7.0 = E_3
\end{aligned}
$$

- En esta parte se emplea la estadística ji–cuadrado dada por la expresión (2.15b) para verificar si las diferencias entre las frecuencias de enfermos observadas (n_i) en las zonas y las esperadas (E_i) se presentaron solo de manera casual. Como la ocurrencia del evento "enfermo" está asociada a una distribución de Poisson siempre que su distribución espacial sea aleatoria, se debe constatar si existe una diferencia estadísticamente significativa entre las frecuencias observadas y las frecuencias esperadas de acuerdo con las probabilidades de Poisson. En la tabla 8.1 se reúnen las frecuencias observadas y esperadas. Con los datos de la tabla 8.1, la es-

Tabla 8.1: Frecuencias observadas y esperadas.

	No. de enfermos			
Frecuencia	0	1	2	3
Observadas (n_i)	41	25	20	14
Esperadas (E_i)	34.3	36.7	19.6	7.0

tadística ji–cuadrado se evalúa de esta manera,

$$\begin{aligned}\chi_0^2 &= \sum_{i=0}^{3} \frac{(n_i - \mathrm{E}_i)^2}{\mathrm{E}_i} \\ &= \frac{(41-34.3)^2}{34.3} + \frac{(25-36.7)^2}{36.7} + \frac{(20-19.6)^2}{19.6} + \frac{(14-7.0)^2}{7.0} \\ &= 12.047\end{aligned}$$

- A la tabla de dimensión 2×4 le corresponden $(2-1) \times (4-1) = 3$ grados de libertad; de donde resultan 2 grados de libertad. Además, como se tuvo que estimar el parámetro μ se pierde otro grado de libertad. El valor del percentil 99 para la ji–cuadrado con 2 grados de libertad es 9.21 (tabla A.2). Dado que el valor de $\chi_0^2 = 12.047$ es mayor que el valor crítico 9.21, se rechaza la hipótesis nula de "*ninguna diferencia*" entre las frecuencias observadas y esperadas, es decir, no se observa un ajuste significativo de los datos a una distribución de Poisson. Como las frecuencias esperadas surgen de suponer una distribución de Poisson, se concluye, según los datos, que la enfermedad no se presenta de manera aleatoria en el área de estudio; en consecuencia, debe indagarse por los posibles factores asociables con su presencia en los diferentes sitios.

8.3 Modelo de regresión tipo Poisson

Se ha insistido que el modelo de Poisson se emplea frecuentemente cuando la variable aleatoria X representa el número de ocurrencias de un evento en un intervalo (de tiempo, espacio o volumen). Una aplicación consiste en verificar si la variable tipo Poisson puede ser explicada por

otras variables. Por ejemplo, el número de defectos dentales en un individuo como una función de la edad, estrato socioeconómico, género, y tipo de pasta dental que emplea para el respectivo aseo.

El modelo de regresión tipo Poisson expresa la media de la variable como una función de algunas variables explicativas $X_1, X_2, \ldots, X_p$, además del tamaño de la unidad sobre la cual se hace el conteo de interés. Por ejemplo, si Y es el número de bacterias en una solución, entonces el *tamaño* es el volumen de la solución; si Y es el número de enfermos, entonces el tamaño corresponde a la magnitud del área de estudio; en el caso de Y, número de piezas dentales defectuosas, el tamaño es el número total de dientes del respectivo individuo.

8.3.1 Modelo de regresión simple

La regresión simple corresponde a un modelo con una sola variable explicativa. En este contexto, se asume que la variable dependiente (o respuesta) Y sigue una distribución de Poisson; los datos del tipo (y_i, x_i) se consiguen desde n unidades de observación. Cuando los eventos de interés ocurren sobre el tiempo, el espacio, o sobre algún otro índice de tamaño, es pertinente modelar la tasa en la cual ocurren tales eventos. Cuando una variable de conteo Y tiene un índice igual a N, la tasa muestral de resultados es Y/N. El valor esperado es μ/N. Para la unidad de observación $i = 1, 2, \ldots, n$, sea N_i el índice relacionado con el tamaño de esta unidad. En el modelo de regresión de Poisson se asume que la relación[1] entre la media de Y y la covariable X se describe por

$$\begin{aligned} \mathrm{E}(Y_i) = \mu_i &= N_i \lambda(x_i) \\ &= N_i \exp[\beta_0 + \beta_1 x_i], \end{aligned} \tag{8.3}$$

donde a $\lambda(x_i) = \mu_i/N_i$ se le denomina el *riesgo* de la i-ésima unidad de observación. A la cantidad $\lambda(x) = \mu/N$ se le llama *tasa de incidencia*, la cual mide la rapidez con que se desarrolla un evento (enfermedad) recientemente considerado (diagnosticado).

El modelo dado en (8.3) es equivalente a

$$\ln\Big(\mu_i/N_i\Big) = \beta_0 + \beta_1 x_i \tag{8.3a}$$

[1]Corresponde a la función de enlace en un modelo lineal generalizado.

o también a:

$$\ln \mu_i = \ln N_i + \beta_0 + \beta_1 x_i \tag{8.3b}$$

En el esquema del modelo lineal generalizado a la cantidad $\ln N_i$ se le llama "*offset*" (de "compensación"), lo cual significa que es una variable cuantitativa cuyo coeficiente de regresión es igual a 1.

Para el modelo presentado en (8.3) (8.3a) o en (8.3b), el número esperado de resultados del evento en consideración satisface

$$\mu = N \exp[\beta_0 + \beta_1 x] = N e^{\beta_0} e^{\beta_1 x} \tag{8.4}$$

Esta media es proporcional al índice de tamaño N, con constante de proporcionalidad que depende del valor de la variable explicativa. Por ejemplo, si se duplica el tamaño de la población también se duplicará el número esperado de eventos.

La estimación de los β se hace vía máxima verosimilitud (sección 1.5.1). Así, bajo el supuesto de que los Y_i son valores de una variable con distribución de Poisson, la función de verosimilitud (de probabilidad) es dada por

$$L(y;\beta) = \prod_{i=1}^{n} \left\{ \frac{[N_i \lambda(x_i)]^{y_i} \exp[N_i \lambda(x_i)]}{y_i!} \right\} \tag{8.5}$$

que equivale, aplicando logaritmo natural en los dos miembros de la expresión, a

$$\ln L(y;\beta) = \sum_{i=1}^{n} \{ y_i \ln N_i - \ln(y_i!) + y_i[\beta_0 + \beta_1 x_i] - N_i \exp[\beta_0 + \beta_1 x_i] \} \tag{8.6}$$

Desde esta última expresión se obtienen los valores de β_0 y β_1, que la maximizan. Estos corresponden a los respectivos estimadores de máxima verosimilitud.

Mediante el procedimiento GENMOD (GENeralized linear MODels) del paquete SAS se hacen los respectivos cálculos, con los cuales se consigue una estimación de estos parámetros, de acuerdo con un conjunto de datos específico.

Medida de asociación

Considere el caso de una variable explicativa dicotómica X que representa una exposición (con 1=expuesto, 0=no expuesto). Entonces se

tiene lo siguiente:

1. Si la i-ésima unidad de observación es expuesta, entonces el logaritmo en la función de riesgo es

$$\ln \lambda_i(\text{expuesto}) = \beta_0 + \beta_1 + \ln N_i,$$

2. Si la i-ésima unidad de observación no es expuesta, entonces

$$\ln \lambda_i(\text{no expuesto}) = \beta_0 + \ln N_i,$$

que es igual a

$$\frac{\lambda_i(\text{expuesto})}{\lambda_i(\text{no expuesto})} = e^{\beta_1}. \tag{8.7a}$$

La cantidad expresada en (8.7a), como se trata en la sección 2.5.5, corresponde al *riesgo relativo* asociado a la exposición.

De manera similar, en el caso de una variable explicativa continua X y cualquier valor x de X,

$$\ln \lambda_i(X = x) = \beta_0 + \beta_1 x + \ln N_i$$
$$\ln \lambda_i(X = x + 1) = \beta_0 + \beta_1(x + 1) + \ln N_i$$

entonces

$$RR = \frac{\lambda_i(X = x + 1)}{\lambda_i(X = x)} = e^{\beta_1}. \tag{8.7b}$$

La expresión (8.7b) representa el riesgo asociado al incremento de una unidad de la variable X.

Considere el número (n_i) de casos nuevos reportados con melanomas en una región. Los totales N_i son los tamaños de las poblaciones que se estiman en riesgo, los cuales pueden representar conteos de personas o conteos de unidades expuestas al riesgo. El interés se dirige a verificar si las tasas n_i/N_i, las cuales son densidades de incidencia, varían a través de dos grupos de personas: quienes trabajan al aire libre o cielo abierto (alibre) y las que trabajan bajo techo (techo). Se recolectaron datos sobre 12 zonas geográficas. La tabla 8.2 contiene los datos. La siguiente es la sintaxis del procedimiento GENMOD del SAS con la cual se hacen los cálculos para la estimación de los parámetros:

Tabla 8.2: Casos nuevos de melanomas.

Zona	Trabajo	Casos (n_i)	Tamaño (N_i)
1	alibre	61	12312
2	alibre	76	15645
3	techo	98	25650
4	techo	104	35650
5	techo	68	19580
6	techo	81	26850
7	alibre	27	8260
8	alibre	45	12340
9	techo	75	21790
10	techo	63	13600
11	techo	52	16580
12	techo	38	6850

```
DATA melanoma;
INPUT indiv trabajo$ casos tamano;
ln=log(tamaño);
CARDS;

(datos)
;
PROC  GENMOD  data=melanoma;
CLASS  trabajo ;
MODEL  casos=edad  / dist=poisson
link=  log offset= ln;
RUN;
```

Los resultados de las estimaciones se muestran en la tabla 8.3. Los

Tabla 8.3: Regresión ajustada a los casos con melanomas.

Parámetro	GL	Estimación	Error est.	ji–cuad.	p–valor
Intercepto	1	−5.6617	0.0416	18560.074	0.0001
Trabajo (alibre)	1	0.2136	0.0807	7.0058	0.0081

resultados de la tabla 8.3 indican que la relación entre el número de casos de melanomas reportados y la exposición a los rayos solares, por trabajar a cielo abierto, es significativa (puesto que p-valor= 0.0081); el riesgo relativo asociado con trabajar al aire libre, de acuerdo con (8.7a), es:

$$\exp[0.2136] = 1.2381.$$

Esto quiere decir que las personas que trabajan expuestas al sol tienen un riesgo del 24% más de contraer melanomas que quienes trabajan bajo techo.

8.3.2 Modelo de regresión múltiple

La variable respuesta de interés puede verse afectada por la influencia de varios factores; esta situación se considera mediante la inclusión de tales factores en el modelo de regresión. De esta manera se postula un modelo de regresión de Poisson múltiple, en el cual las variables explicativas se combinan de manera lineal. Las variables explicativas o independientes pueden ser dicotómicas, politómicas o continuas; las variables categóricas se representan mediante variables ficticias (sección 5.5).

Si con las variables $X_1, X_2, \ldots, X_p$ se pretende explicar y predecir la variable Y mediante un modelo de regresión de Poisson múltiple, un modelo adecuado puede ser el siguiente:

$$\begin{aligned} \mathrm{E}(Y_i) = \mu_i &= N_i \lambda(x_{i1}, x_{i2}, \ldots, x_{ip}) \\ &= N_i \exp[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}] \\ &= N_i \exp[\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}], \end{aligned} \tag{8.8}$$

donde Y es la variable dependiente con distribución de Poisson y $\lambda(x_{i1}, x_{i2}, \ldots, x_{ip})$ es el riesgo al que se expondría la i-ésima unidad de observación. El modelo contenido en (8.8) es equivalente a

$$\ln\Big(\mu_i/N_i\Big) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{8.8a}$$

o también a

$$\ln \mu_i = \ln N_i + [\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}] \tag{8.8b}$$

La estimación de los β se hace de manera análoga (mediante máxima verosimilitud) a la desarrollada para el caso de una variable. Mediante el procedimiento GENMOD del paquete SAS, se obtienen los cálculos para las estimaciones asociadas a un conjunto de parámetros, de acuerdo con los datos disponibles.

Como en la regresión lineal simple, $\exp[\beta_i]$ sirve para calcular:

(i) El riesgo relativo asociado a una exposición si X_i es dicotómica (expuesto es $X_i = 1$ y no expuesto es $X_i = 0$).

(ii) El riesgo relativo por el incremento en una unidad si X_i es continua ($X_i = x + 1$ frente a $X_i = x$).

(iii) Un intervalo de confianza, por ejemplo del 95%, para el riesgo relativo anterior viene dado por

$$\exp\{\widehat{\beta}_i \mp 1.96\,\mathrm{ee}(\widehat{\beta}_i)\} = \left[\exp\{\widehat{\beta}_i - 1.96\,\mathrm{ee}(\widehat{\beta}_i)\}; \exp\{\widehat{\beta}_i + 1.96\,\mathrm{ee}(\widehat{\beta}_i)\}\right] \tag{8.9}$$

Este intervalo se interpreta de acuerdo con la contenencia o no del 1. Si el 1 está contenido se puede afirmar que el riesgo relativo es igual sea el caso (i) o el (ii); si ambos extremos del intervalo son menores que 1, hay más riesgo en la exposición (o en $X = x + 1$) que en el otro caso; y recíprocamente, si ambos extremos del intervalo son superiores a 1, hay más riesgo de la no exposición (o en $X = x$) que en el otro caso.

Como se acostumbra en el análisis de regresión, los datos deben examinarse para asegurar que se satisfacen los requerimientos del modelo. Por ejemplo, si algunas de las variables están altamente correlacionadas, entonces con solo algunas de ellas se pueden obtener predicciones tan buenas como con todas las variables. El uso de la interacción entre variables de la forma $X_i X_{i'}$, o de potencias del tipo X_i^2, pueden coadyuvar a un ajuste adecuado del modelo a los datos y, en consecuencia, a la lectura e interpretación del problema en estudio.

Verificación de hipótesis sobre los parámetros

Se trata de responder a la inquietud sobre la *contribución de cada variable* en la predicción de la respuesta con distribución Poisson. Se tienen tres procedimientos que responden a este interrogante:

(i) Una verificación global del ajuste; es decir, dar cuenta si las variables en conjunto contribuyen significativamente a la predicción de la respuesta.

(ii) Una inspección particular de cada variable o factor; para advertir sobre su importancia con relación a las demás, en la predicción de la respuesta.

(iii) Una verificación de un grupo de variables; para observar su contribución, respecto a las demás variables independientes, en la predicción de la respuesta.

Verificación global

La inquietud (i) apunta a una verificación general sobre todas las p variables explicativas incorporadas al modelo; esto se aborda mediante la verificación de la hipótesis nula: "ninguna de las p variables independientes explican la variación de la respuesta aparte del índice de tamaño de la unidad de observación". También se expresa así:

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

Esta hipótesis se verifica mediante la razón de verosimilitud ji–cuadrado con p grados de libertad

$$\chi^2 = 2[L_p - L_0] \tag{8.10}$$

donde L_p es el logaritmo del valor de la función de verosimilitud para el modelo que contiene todas las p variables y L_0 es el logaritmo de la función de verosimilitud para el modelo que contiene solo el intercepto (bajo H_0). Con el procedimiento GENMOD del SAS, se calculan estos valores.

Verificación sobre una variable específica

El procedimiento (ii) último se traduce en verificar si la adición de alguna variable independiente contribuye (o agrega) significativamente a la predicción del modelo con relación a las demás variables ya incorporadas al modelo. La hipótesis nula para esta verificación se puede escribir como: "la variable X_i no adiciona valor alguno a la predicción

de la respuesta dado que las demás variables han sido incluidas en el modelo"; de otra manera

$$H_0 : \ \beta_i = 0.$$

Esta hipótesis se verifica mediante

$$Z_i = \frac{\widehat{\beta}_i}{\text{ee}(\widehat{\beta}_i)} \tag{8.11}$$

donde $\text{ee}(\widehat{\beta}_i)$ es el error estándar de $\widehat{\beta}_i$. Z_i es la estadística de Wald. Se rechaza H_0 para valores de Z_i mayores, en valor absoluto, que el percentil $(1 - \alpha/2)$ de la distribución normal estándar.

Contribución de un grupo de variables

Este procedimiento es una generalización del anterior, pues se trata de verificar si dos o más variables contribuyen significativamente a la predicción de la respuesta por encima de las demás. Esto equivale a verificar la siguiente hipótesis nula

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_k = 0, \text{con } k < p.$$

Para verificar esta hipótesis se desarrolla la estadística de razón de verosimilitud ji–cuadrado con k grados de libertad

$$\begin{aligned} \chi^2 = 2[&\ln L(\widehat{\beta}; \text{en todas las } X) \\ &- \ln L(\widehat{\beta}; \text{en las } X \text{ sin las } k \ X \text{ de interés})] \end{aligned} \tag{8.12}$$

Este procedimiento es útil cuando se quiere corroborar si un grupo de variables (demográficas, culturales, antropomórficas, fisiológicas, etc.) es conjuntamente importante en la predicción de la respuesta.

Considere la variable dependiente que cuenta el número de casos con cáncer de piel (Le 1998, 221). Los datos fueron obtenidos en dos áreas metropolitanas. La población de cada área fue dividida en ocho grupos de edad. La tabla 8.4 recoge los datos del estudio. En este ejemplo se involucran dos variables explicativas: la edad y la localización. Ambas variables se han considerado como categóricas; mediante variables ficticias o *dummy* (con 85+ como base) se pueden representar los ocho

Tabla 8.4: Datos sobre cáncer en la piel.

	Área metropolitana			
	Área 1		Área 2	
Grupo etáreo	Casos	Pobla.	Casos	Pobla.
15-24	1	172675	4	181343
25-34	16	123065	38	146207
35-44	30	96216	119	121374
45-54	71	92051	221	111353
55-64	102	72159	259	83004
65-74	130	54722	310	55932
75-84	133	32185	226	29007
85+	40	8328	65	7538

grupos etáreos y una para localización (con Área 1 como base). La estimación de los parámetros se hace mediante las siguientes instrucciones dentro del procedimiento GENMOD del paquete SAS:

```
DATA  cancer;
INPUT edad area$ casos pobla$;
ln=LOG(pobla);
CARDS;
15-24 A1  1 172675
        .
        .
        .
85+   A2  65 7538
;
PROC  GENMOD  data=cancer;
CLASS  edad  area ;
MODEL  casos=edad area  /
DIST=poisson LINK= log
OFFSET=ln;
RUN;
```

Los resultados de la estimación, los errores estándar, el valor de la estadística ji–cuadrado (el cuadrado de la estadística de Wald, ecuación (8.11) y el p-valor respectivo se muestran en la tabla 8.5.

Tabla 8.5: Estimación del modelo de regresión múltiple con los datos de la tabla 8.4.

Variable		G.L.	Estimac.	E. estándar	ji–cuad.	valor–p
Intercep		1	-4.6750	0.0991	2225.5549	0.0001
AREA	Área 2	1	0.8043	0.0522	237.3441	0.0001
EDAD	15-24	1	-6.1782	0.4577	182.1726	0.0001
EDAD	25-34	1	-3.5480	0.1675	448.7600	0.0001
EDAD	35-44	1	-2.3308	0.1275	334.3631	0.0001
EDAD	45-54	1	-1.5830	0.1138	193.3780	0.0001
EDAD	55-64	1	-1.0909	0.1109	96.7471	0.0001
EDAD	65-74	1	-0.5328	0.1086	24.0581	0.0001
EDAD	75-84	1	-0.1196	0.1109	1.1629	0.2809

Estos indican una tendencia ascendente de la incidencia de este cáncer con la edad. El riesgo relativo de contraer cáncer de piel, asociado con vivir en el Área 2, teniendo como base vivir en en el Área 1, es

$$RR = \exp[0.8043] = 2.235$$

es decir, que en el Área 2 este riesgo se incrementa en más del doble respecto a vivir en el Área 1.

Para considerar el efecto conjunto de las siete variables ficticias que representan la edad se procede a verificar si es significativa la diferencia entre el modelo con las ocho variables (edad y sitio) y el modelo que tiene solo las de sitio.

1. Con todas las variables, el logaritmo de la verosimilitud es 7201.9.

2. Cuando se sacan las siete variables, el logaritmo de la verosimilitud es 5921.1 En consecuencia,

$$\begin{aligned}\chi^2 &= 2[\ln L(\widehat{\beta}; \text{con las 8 variables}) \\ &\qquad - \ln L(\widehat{\beta}; \text{con sólo la variable de localización})] \\ &= 2(7201.9 - 5921.1) \\ &= 2561.6\end{aligned}$$

el cual tiene asociados 7 grados de libertad con un p-valor menor que 0.0001; es decir, que los grupos de edad influyen significativamente más que los sitios. Estos cálculos se logran incluyendo la instrucción *Type3* en el procedimiento GENMOD; la sintaxis es la siguiente:

```
MODEL CASOS=EDAD AREA / DIST=POISSON
LINK=LOG OFFSET=LN TYPE3;
```

Los resultados con relación a la influencia de la edad y del área de residencia en el cáncer de piel se muestran en la tabla siguiente. Se

Tabla 8.6: Razón de verosimilitud.

Factor	GL	ji–cuadrado	p-valor
Edad	7	2561.5700	< 0.0001
Área	1	258.7193	< 0.0001

puede advertir que las dos variables tienen un efecto (a la manera de efectos principales) significativo en el número de casos con cáncer en la piel reportados.

En el análisis tipo 3 (del SAS) se verifica el efecto de un factor (algunas variables) con relación a los demás factores; es decir, se observa la contribución adicional de cada factor en la explicación de la variable respuesta. A veces, puede ser de interés la verificación de un ajuste secuencial. Por ejemplo, cuando se tiene la variable Y con distribución de Poisson y las variables X_1, X_2, y X_3, se quiere ver el efecto de ajustar la variable X_1 sobre Y, el efecto de agregar la variable X_2 al modelo que contiene la variable X_1 y el efecto de agregar X_3 al modelo que contiene las variables X_1 y X_2. Este *análisis secuencial* puede hacerse con el PROC GENMOD del SAS con la opción TYPE1 (con sintaxis similar a la opción TYPE3).

8.4 Procesamiento de datos con R

En esta sección se presenta la forma de ajustar modelos de regresión tipo Poisson, usando la función `glm()`.

Modelo de regresión simple tipo Poisson

Se ilustra cómo ajustar un modelo de regresión simple tipo Poisson, como el descrito por la ecuación 8.3b, considerando los datos de la tabla 8.2.

```
# introducción de los datos
Trabajo<-c("alibre","alibre","techo","techo",
           "techo","techo","alibre","alibre",
           "techo","techo","techo","techo")
# definición de trabajo como un factor con
#techo como nivel de referencia
Trabajo<-relevel(factor(Trabajo),ref="techo")
# columna casos
casos<-c(61,76,98,104,68,81,27,45,75,63,52,38)
# columna tamaño
tamano<-c(12312,15645,25650,35650,19580,26850,
          8260,12340,21790,13600,16580,6850)
# se organizan los datos en un data.frame y se
# crea la columna con el logaritmo de tamaño
tabla8.2<-data.frame(Trabajo,casos,tamano,
                     lt=log(tamano))
# ajuste del modelo
mod8.3b<-glm(casos~offset(lt) + Trabajo, data=tabla8.2,
          family=poisson(link = "log"))
summary(mod8.3b)
```

Modelo de regresión múltiple tipo Poisson

Para el ajuste de un modelo de regresión múltiple tipo Poisson, se usan los datos de la tabla 8.4.

```
# grupos de edad
ge<-c("15-24","25-34","35-44","45-54","55-64","65-74",
      "75-84","85+")
# se define ge como factor y 85+ como su nivel de
# referencia
ge<-relevel(factor(ge),ref="85+")
# columna casos
```

```
casos<-c(1,16,30,71,102,130,133,40,4,38,119,221,259,
         310,226,65)
# columna población
pobla<-c(172675,123065,96216,92051,72159,54722,32185,
         8328,181343,146207,121374,111353,83004,55932,
         29007,7538)
# columna area como un factor
area<-factor( rep(c("A1","A2"),c(8,8)) )
#se organizan los datos en un data.frame
tabla8.4<-data.frame(ge,area,casos,pobla,lp=log(pobla))
# ajuste del modelo
mod8.8b<-glm(casos~offset(lp) + area + ge, data=tabla8.4,
          family=poisson(link = "log"))
# se imprimen los resultados del ajuste
summary(mod8.8b)
```

A continuación se ajusta el modelo que no incluye el efecto de edad y se evalúa si la diferencia en el ajuste es significativa, usando la función `anova()`

```
mod2<-glm(casos~offset(lp) + area, data=tabla8.4,
          family=poisson(link = "log"))
summary(mod2)
anova(mod2,mod8.8b,test="Chisq")
```

La evaluación del efecto de la entrada secuencial de variables a un modelo se evalúa mediante la función anova, de la siguiente manera:

```
anova(mod8.8b,test="Chisq")
```

8.5 Ejercicios

1. Baxter, Coutts & Ross (1980), citado por Dobson (2002), reportan datos sobre el número de pólizas de seguro, `n`, y el número de reclamos, `y`, para vehículos en varias categorías de seguros `CAR`, tabuladas por la edad del tomador de la póliza, `AGE`, el distrito donde vive el tomador de la póliza: `DIST=1` para Londres y otras ciudades grandes, `DIST=0` en otro caso. Los datos se transcriben en la tabla 8.7. El objetivo es estimar el efecto de los factores `CAR`, `AGE` y `DIST` sobre el número de pólizas que hacen reclamo.

   a) Decida entre el modelo tipo Poisson que incluya todos los factores junto con todas las interacciones posibles y el modelo que no incluye los términos de interacción. ¿Qué criterio usó para tomar la decisión?

   b) Discuta la bondad de ajuste del modelo elegido. ¿Qué criterio usó?

   c) Interprete los parámetros del modelo final en términos de razones de promedios.

2. Los datos en la tabla 8.8 son de un estudio famoso conducido por sir Richard Dolly y colegas. En 1951 le enviaron a todos los doctores británicos un breve cuestionario que preguntaba si fumaban tabaco. Desde entonces se recolectó información acerca de su muerte. La tabla 8.8 muestra el número de muertes de doctores hombres debido a enfermedades coronarias 10 años después del estudio. La tabla también contiene el numero de personas–año en el momento del análisis. Las preguntas de interés son:

   a) ¿Es la tasa de muertes más grande entre fumadores que entre los no fumadores? Si es así, ¿qué tanto?

   b) ¿Hay algún efecto diferencial debido a la edad?

3. Se realizó un experimento para determinar el efecto de la radiación gama sobre el número de anormalidades cromosómicas observadas en la células[2]. Los datos se consiguen en el marco de

[2]Purott R. and Reeder E. (1976) The effect of changes in dose rate on the yield of chromosome aberrations in human lymphocytes exposed to gamma radiation. Mutation Research. 35, 437-444.

Tabla 8.7: Número de pólizas de seguros y número de reclamos.

Vehíc.	Edad	y	n	Dist.	Vehíc.	Edad	y	n	Dist.
1	1	65	317	0	1	1	2	20	1
1	2	65	476	0	1	2	5	33	1
1	3	52	486	0	1	3	4	40	1
1	4	310	3259	0	1	4	36	316	1
2	1	98	486	0	2	1	7	31	1
2	2	159	1004	0	2	2	10	81	1
2	3	175	1355	0	2	3	22	122	1
2	4	877	7660	0	2	4	102	724	1
3	1	41	223	0	3	1	5	18	1
3	2	117	539	0	3	2	7	39	1
3	3	137	697	0	3	3	16	68	1
3	4	477	3442	0	3	4	63	344	1
4	1	0	3	0	4	1	0	3	1
4	2	6	16	0	4	2	6	16	1
4	3	8	25	0	4	3	8	25	1
4	4	33	114	0	4	4	33	114	1

Tabla 8.8: Muertes por enfermedades coronarias.

Grupo Edad	Muertes	Personas	Fuma
1	32	52407	Sí
2	104	43248	Sí
3	206	28612	Sí
4	186	12663	Sí
5	102	5317	Sí
1	2	18790	No
2	12	10673	No
3	28	5710	No
4	28	2585	No
5	31	1462	No

datos `dicentric` de la librería `faraway` constan de 27 observaciones con las siguiente 4 variables: `cells`: Número de células en cientos, `ca`: Número de anormalidades cromosómicas, `doseamt`: Cantidad de dosis en Grays, ensayado en tres niveles: 1, 2.5, 5,

`doserate`: tasa de dosis en Grays/hora, ensayado en los niveles: 0.1, 0.25, 0.5, 1, 1.5, 2, 2.5, 3, 4. Ajuste el modelo $ca = \log(cells) + \log(doserate) * doseamt$, donde `doseamt` se toma como un factor. Discuta la bondad de ajuste del modelo e interprete los valores de los parámetros estimados.

Capítulo 9

Métodos para datos emparejados

9.1 Introducción

En el estudio de casos (personas con una condición, por ejemplo, enfermedad) y controles (personas sin la condición), los *casos* son colocados en las mismas condiciones[1] (equiparados, emparejados o pareados) que los *controles* sobre la base de variables que se piensan son potencialmente confusoras (extrañas) de la respuesta, como la edad, la raza, el género, entre otras.

Ejemplos de emparejamiento u homogeneización son los siguientes:

- Asignarle a un individuo *caso* un individuo *control* que tenga las mismas características demográficas; de manera que el interés se dirija a determinar si hay diferencia entre la exposición del control al factor de riesgo y la exposición del caso al mismo factor de riesgo.

- Medidas registradas sobre el ojo izquierdo y derecho, respectivamente, en varios individuos; en general, mediciones realizadas en puntos corporales equivalentes.

[1] En diseño experimental se habla de conformar bloques.

- Medidas de un mismo individuo en dos puntos distintos de tiempo (medidas repetidas).
- Otro tipo de emparejamiento en personas son los vecinos y las familias, con quienes, se espera, está bajo condiciones genéticas, etnográficas, demográficas y ambientales similares.

La idea es colocar las unidades de observación en condiciones suficientemente homogéneas, de manera que la variabilidad en las respuestas observadas pueda ser atribuida a la exposición o no ante el factor de riesgo (el cual puede ser un tratamiento).

Algunas de las estrategias para controlar efectos de confusión son el empleo de la aleatorización y la estratificación (bloqueo) en la etapa del planeamiento o diseño, lo mismo que la realización de algunos ajustes para la etapa del análisis estadístico de los datos. Una forma usual de estratificación se describe en el caso de diseños en bloques (7.5), donde cada individuo–caso se empareja con uno o más individuos–control, seleccionados con características similares.

Algunos tipos de emparejamiento mediante casos y controles son:

- Un equiparamiento $1:1$ consiste en estudios en los cuales cada caso asocia a un control.
- Un equiparamiento $1:m_i$ consiste en estudios en los cuales al $i-$ésimo caso se le asocian m_i controles.
- Un equiparamiento $m:n$ consiste en estudios en los cuales m casos se asocian a n controles; usualmente m y n varían entre 1 y 5.

Los diseños pareados tienen varias ventajas. Permiten controlar variables de confusión difíciles de medir directamente, y por tanto, difíciles de ajustar en la etapa de análisis de los datos. El emparejamiento provee también un control adecuado de la confusión, superior al que se hace mediante la regresión, porque este no requiere supuestos como la forma funcional del modelo, la cual sí es demandada en el análisis de regresión. Una desventaja es que para un individuo caso con características poco comunes, puede resultar difícil encontrar individuos con el mismo perfil. Además, cuando los casos y los controles son emparejados sobre alguna característica específica, la influencia de tal característica puede que no

llegue a estudiarse ampliamente; es decir, que la variable de confusión sea opacada por la misma estrategia de control. Además, la muestra pareada de casos y controles usualmente no representa una población determinada, lo cual restringe las posibilidades de inferencia o generalización.

9.2 Medidas de concordancia o acuerdo

Una práctica seguida por muchos pacientes, incluso sin intención deliberada estadística, es comparar los resultados que sobre una misma enfermedad reportan dos observadores diferentes (por ejemplo, laboratorios). Otro escenario para comparar la respuesta que un individuo muestra ante una situación específica es el tiempo; casos en los que a un mismo grupo de individuos se les observa en dos momentos diferentes. En situaciones como estas la atención se dirige a dar cuenta sobre la concordancia o no de las respuestas. En esta sección se exhiben algunas medidas con las cuales se pueda advertir acerca del acuerdo o desacuerdo de los resultados en dos muestras, cuando cada muestra tiene los mismos sujetos o cuando a un sujeto de una muestra se "empareja" con uno determinado de la otra muestra. Las técnicas estadísticas con las que se explora este tipo de datos deben considerar la *dependencia* estadística existente entre las dos muestras.

Para orientar la presentación, admita que se tienen dos observadores, cada uno de los cuales asigna independientemente a cada uno de n individuos en una de dos (aunque pueden ser más) categorías (A o B). La tabla 9.1 muestra los posibles resultados. Las probabilidades calculadas

Tabla 9.1: Concordancia entre dos observadores.

	Observador 2		
Observador 1	Categoría A	Categoría B	Total
Categoría A	n_{11}	n_{12}	$n_{1.}$
Categoría B	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n = n_{11} + n_{12} + n_{21} + n_{22}$

con los datos anteriores se consignan en la tabla 9.2. Con las frecuencias anteriores se define:

Tabla 9.2: Probabilidades de concordancia entre dos observadores

	Observador 2		
Observador 1	Categoría A	Categoría B	Total
Categoría A	p_{11}	p_{12}	$p_{1.}$
Categoría B	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	$1.0 = p_{11} + p_{12} + p_{21} + p_{22}$

1. *La proporción global de concordancia*:

$$C = \frac{n_{11} + n_{22}}{n} \tag{9.1}$$

2. *La proporción específica de concordancia*:

$$\begin{aligned} C_A &= \frac{2n_{11}}{2n_{11} + n_{12} + n_{21}}, \text{ para la categoría } A \\ C_B &= \frac{2n_{22}}{2n_{22} + n_{12} + n_{21}}, \text{ para la categoría } B \end{aligned} \tag{9.2}$$

La diferencia entre asociación y concordancia es: para que exista una asociación perfecta se requiere únicamente que se pueda predecir la categoría en una respuesta desde la categoría de la otra respuesta; para que dos respuestas tengan concordancia perfecta, deben caer en categorías idénticas. No obstante, las proporciones de concordancia (global o específica) no miden acuerdo.

Se sugiere comparar la concordancia global (proporciones de la tabla 9.2),

$$\theta_1 = \sum_i p_{ii} \tag{9.3}$$

con los cambios de concordancia,

$$\theta_2 = \sum_i p_{i.} p_{.i} \tag{9.4}$$

que ocurren si la variable fila es independiente de la variable columna. Esto se justifica por el hecho de que la independencia entre la variable de filas y la variable de columnas, se tiene cuando el producto de las probabilidades marginales es igual a la probabilidad de la celda (probabilidad conjunta). Una medida de acuerdo es la llamada estadística *kappa*:

$$\kappa = \frac{\theta_1 - \theta_2}{1 - \theta_2} = \frac{\sum_i p_{ii} - \sum_i p_{i.}p_{.i}}{1 - \sum_i p_{i.}p_{.i}} \tag{9.5}$$

donde el subíndice i recorre las categorías o modalidades de la variable.

Esta medida es apropiada para escalas nominales. El numerador compara la probabilidad de acuerdo que se espera bajo la hipótesis de independencia. El denominador reemplaza a $\sum_i p_{ii}$ por su máximo posible 1.0, que corresponde al "acuerdo perfecto". La estadística *kappa* varía entre 0.0 y 1.0, correspondiendo a 1.0 un acuerdo perfecto.

Para el caso de dos categorías (tablas 2×2) la estadística *kappa* es equivalente a

$$\kappa = \frac{2[n_{11}n_{22} - n_{12}n_{21}]}{n_{1.}n_{.2} + n_{.1}n_{2.}} \tag{9.6}$$

El error estándar (ee(;)) es dado por

$$\text{ee}(\kappa) = \frac{2\sqrt{[n_{11}n_{22} - n_{12}n_{21}]/2}}{n^2 - [n_{1.}n_{.1} + n_{2.}n_{.2}]} \tag{9.7}$$

A continuación se muestra qué sugiere la estadística *kappa* en investigación clínica, en relación con los valores que tome:

$\kappa > 0.75$	excelente acuerdo
$0.40 \leq \kappa \leq 0.75$	buen acuerdo
$0 \leq \kappa < 0.40$	pobre acuerdo

Existe una gran controversia sobre la utilidad del *kappa*, principalmente porque su valor depende fuertemente de las distribuciones marginales. No obstante, se puede considerar como una primera señal en un esquema descriptivo.

Dentro del procedimiento FREQ del paquete SAS se ofrece el cálculo de la estadística *kappa* y de su valor p.

Se pidió a un grupo de neurólogos de hospitales diferentes que clasificaran a una serie de pacientes en una de cuatro categorías de respuesta respecto a un diagnóstico de esclerosis múltiple[2]:

(1) Caso seguro (2) Caso probable (razón 3:1)
(3) Caso dudoso (razón 1:1) (4) Caso improbable

[2]Adaptado de Ato & López (1996, 346–348).

Los datos que dan cuenta de la clasificación de los dos neurólogos se disponen en la tabla 9.3; entre paréntesis están las respectivas probabilidades. Los cálculos de la estadística *kappa*, de acuerdo con la expresión

Tabla 9.3: Diagnóstico de dos neurólogos.

	Neurólogo 2				
Neurólogo 1	Seguro	Muy probable	Dudoso	Improbable	Total
Seguro	38 (38/149)	5 (5/149)	0 (0/149)	1 (1/149)	44 (44/149)
Muy Probable	33 (33/149)	11 (11/149)	3 (3/149)	0 (0/149)	47 (47/149)
Dudoso	10 (10/149)	14 (14/149)	5 (5/149)	6 (6/149)	35 (35/149)
Improbable	3 (3/149)	7 (7/149)	3 (3/149)	10 (10/149)	23 (23/149)
Total	84 (84/149)	37 (37/149)	11 (11/149)	17 (17/149)	149 (149/149)

(9.5), son los siguientes:

$$
\begin{aligned}
\kappa &= \frac{\sum_i p_{ii} - \sum_i p_{i.}p_{.i}}{1 - \sum_i p_{i.}p_{.i}} \\
&= \frac{\left[\frac{38}{149}+\frac{11}{149}+\frac{5}{149}+\frac{10}{149}\right] - \left[\frac{44}{149}\times\frac{84}{149}+\frac{47}{149}\times\frac{37}{149}+\frac{35}{149}\times\frac{11}{149}+\frac{23}{149}\times\frac{17}{149}\right]}{1-\left[\frac{44}{149}\times\frac{84}{149}+\frac{47}{149}\times\frac{37}{149}+\frac{35}{149}\times\frac{11}{149}+\frac{23}{149}\times\frac{17}{149}\right]} \\
&= \frac{\frac{64}{149} - \frac{6211}{22201}}{1-\frac{6211}{22201}} \\
&= \frac{0.4295-0.2798}{1-0.2798} \\
&= \frac{0.1497}{0.7202} \\
&= 0.2079
\end{aligned}
$$

Por las indicaciones presentadas ($\kappa < 0.40$), se advierte que hay un pobre o débil acuerdo entre los diagnósticos dados por los dos neurólogos. Esta situación debe invitar a una indagación más compleja sobre el problema o a la realización de un panel de expertos, por ejemplo.

9.3 Estudios emparejados caso-control

Considere un estudio caso-control y suponga que cada individuo es clasificado como *expuesto* o *no expuesto* a cierto factor, y, por ejemplo, padecer o no padecer cierta enfermedad. La población puede clasificarse en una tabla 2×2, donde las celdas son las proporciones de la modalidad fila y columna respectiva con relación al total. La tabla 9.4 describe, de esta manera, la población. La asociación entre el factor y la enfermedad

Tabla 9.4: Proporciones factor × enfermedad.

	Enfermedad		
Factor	+	−	Total
+	P_1	P_3	$P_1 + P_3$
−	P_2	P_4	$P_2 + P_4$
Total	$P_1 + P_2$	$P_3 + P_4$	1.0

puede medirse por la razón de riesgos o riesgo relativo (RR) de padecer la enfermedad (+) con o sin el factor de riesgo

$$\begin{aligned} RR &= \frac{P_1}{P_1 + P_3} \div \frac{P_2}{P_2 + P_4} \\ &= \frac{P_1(P_2 + P_4)}{P_2(P_1 + P_3)}. \end{aligned} \tag{9.8}$$

En la mayoría de los casos, P_1 es pequeño comparado con P_3; P_2 es pequeño comparado con P_4, de manera que el riesgo relativo es aproximadamente igual a la razón de *odds* (expresiones 2.46a, 2.46b, 2.46c o 2.46d) de padecer la enfermedad:

$$\begin{aligned} RO &= \frac{P_1 P_4}{P_2 P_3} \\ &= \frac{P_1/P_3}{P_2/P_4}, \end{aligned} \tag{9.9}$$

o la razón de *odds* de estar expuesto al factor

$$RO = \frac{P_1/P_2}{P_3/P_4}. \tag{9.10}$$

Como una estrategia para controlar factores de confusión los casos son emparejados, uno a uno, con un conjunto de controles escogidos con valores similares en las variables de confusión relevantes (nuevamente, una especie de "bloqueo"). El caso más sencillo se tiene para un factor de exposición dicotómico (por ejemplo, "fumar" frente a "no fumar"). Este tipo de datos se condensan en la tabla 9.5 Así por ejemplo, n_{10} señala los pares donde el caso es expuesto, pero su pareja control no es expuesta. El modelo estadístico más apropiado para hacer inferencias

Tabla 9.5: Frecuencias caso × control.

	Caso		
Control	+	-	Total
+	n_{11}	n_{01}	$n_{11}+n_{01}$
-	n_{10}	n_{00}	$n_{10}+n_{00}$
Total	$n_{11}+n_{10}$	$n_{01}+n_{00}$	n

sobre la razón de *odds* es la probabilidad condicional del número de casos expuestos entre los pares discordantes. Dado que $(n_{10}+n_{01})$ se fija previamente, se puede considerar que n_{10} tiene distribución binomial $B(n_{10}+n_{01},\pi)$ (sección 1.4.2), donde

$$\pi=\frac{RO}{1+RO};\ \text{en consecuencia, } 1-\pi=\frac{1}{1+RO}$$

La función de verosimilitud para la muestra de tamaño $n_{10}+n_{01}$ es

$$\pi^{n_{10}}(1-\pi)^{n_{01}}=\Big[\frac{RO}{1+RO}\Big]^{n_{10}}\Big[\frac{1}{1+RO}\Big]^{n_{01}}$$

De donde el estimador para la razón de *odds* es

$$\widehat{RO}=\frac{n_{10}}{n_{01}} \tag{9.11}$$

con

$$\widehat{\text{var}}(\widehat{RO})=\frac{n_{10}(n_{10}+n_{01})}{n_{01}^3} \tag{9.12}$$

Una aplicación de (9.11) y (9.12) es la construcción de intervalos de confianza del 95% para estimar la razón de *odds*; estos se calculan mediante

$$\widehat{RO}\mp(1.96)\sqrt{\widehat{\text{var}}(\widehat{RO})} \tag{9.13}$$

Equivalentemente, se verifica la hipótesis nula de efecto del riesgo mediante la estadística

$$z=\frac{(n_{10}-n_{01})}{\sqrt{n_{10}+n_{01}}}\sim n(0,1) \tag{9.14}$$

Cuyo cuadrado es equivalente a la estadística de McNemar (sección 2.5.4)

$$z^2=\chi^2_M=\frac{(n_{10}-n_{01})^2}{n_{10}+n_{01}}, \tag{9.15}$$

la cual tiene una distribución ji–cuadrado con un grado de libertad.

Un procedimiento alterno es el Cochran-Mantel-Haenszel (sección 2.5.8), el cual produce el mismo estimador para la razón de *odds*, es decir,

$$\widehat{RO}_{CMH} = \widehat{RO} = \frac{n_{10}}{n_{01}} \tag{9.16}$$

Con muestras de tamaño grande se estima RO mediante un intervalo del 95% de confianza, como sigue:

$$\frac{n_{10}}{n_{01}} \exp\left[\mp(1.96)\sqrt{\frac{1}{n_{10}} + \frac{1}{n_{01}}}\right] \tag{9.17}$$

Un estudio especial de casos y controles es en el que cada caso se acompaña de un control. Para una respuesta dicotómica Y, cada caso $(Y = 1)$ se parea con un control $(Y = 0)$ de acuerdo con cierto(s) criterio(s) X que podría(n) afectar la respuesta. Cada par de sujetos es medido con relación a la variable de predicción X, y se analiza la posible asociación entre las variables X y Y. A continuación se muestra un ejemplo con este tipo de estudios[3].

Los datos de la tabla 9.6 corresponden al emparejamiento simple de casos y controles. En un estudio sobre la agudeza de infartos en el miocardio (IM), 144 víctimas de IM se emparejaron, de acuerdo con la edad y el género, con 144 personas libres de enfermedades cardiacas. A estas personas se les indagó acerca de si habían sido diagnosticados como diabéticos o no ($X = 0$, es no, y $X = 1$, es sí). Una estimación de

Tabla 9.6: Diagnóstico previo de diabetes para MI. Pares caso-control.

	Casos IM		
Controles IM	Diabetes	No diabetes	Total
Diabetes	9	16	25
No diabetes	37	82	119
Total	46	98	144

la razón de *odds* por (9.11) es

$$\begin{aligned}\widehat{RO} &= \frac{n_{10}}{n_{01}} = \frac{37}{16} \\ &= 2.3\end{aligned}$$

[3] Agresti (1996, 232).

de donde se puede afirmar que la probabilidad de que un individuo diabético padezca infarto en el miocardio es 2.3 veces mayor que la de un individuo no diabético.

9.4 Regresión logística condicional

Se recuerda que el término "emparejamiento" se refiere a parear uno o más controles con cada caso con base en su similaridad (bloques o estratos) respecto de algunas variables criterio. Como uno de los propósitos centrales del estudio es indagar sobre la posible asociación entre el evento de interés, generalmente una enfermedad, y un conjunto de variables o factores empleadas para emparejar las casos con los controles. La *regresión logística condicional* es la técnica de modelamiento apropiada para estos requerimientos.

La regresión logística condicional se justifica también por la insuficiencia de los tamaños de muestra en cada uno de los estratos que definen las variables de emparejamiento, particularmente si el número de estratos es alto. La estratificación se incorpora en la estimación del modelo logístico mediante la *función de verosimilitud condicional*. De esta manera, la *función de verosimilitud condicional* ayuda a resolver el caso de que un número de parámetros mayor que el número de observaciones disponible en los datos lleva a problemas de estimación.

9.4.1 Regresión logística simple

Se considera el modelo en el cual se tiene en cuenta una sola variable regresora para predecir la variable dicotómica de interés. Se desarrolla, de manera esquemática, esta técnica con la estructura caso-control de la forma un caso emparejado con m_i controles; note que el número de controles m_i puede variar de un caso a otro. En la tabla 9.7 se muestra este tipo de emparejamiento. Así, el valor de la variable regresora para cada *caso* i se escribe x_i y el valor de la variable regresora para el j-ésimo *control* emparejado con el caso i se escribe x_{ij}. De esta forma, para el i-ésimo emparejamiento, la probabilidad condicional del resultado observado (que al individuo con valor x_i le ocurra el evento) dado que se tiene un caso (a quien no le ocurre el evento), por cada grupo de

Tabla 9.7: Emparejamiento $1 : m_i$.

Caso	Control
i	1
	2
	$\vdots$
	m_i

emparejamiento (tabla 9.7), es

$$\frac{\exp(\beta x_i)}{\exp(\beta x_i) + \sum_{j=1}^{m_i} \exp(\beta x_{ij})}. \tag{9.18}$$

La función de verosimilitud condicional para todos los grupos de emparejamiento, suponiendo que la muestra consta de n casos, es

$$L = \prod_{i=1}^{n} \frac{\exp(\beta x_i)}{\exp(\beta x_i) + \sum_{j=1}^{m_i} \exp(\beta x_{ij})}. \tag{9.19}$$

A partir de esta expresión y mediante algunos procedimientos de cálculo se obtienen los estimadores de máxima verosimilitud para los β.

Para el caso en el que $m_i = 1$ (emparejamiento $1 : 1$), y la variable regresora X, donde $X = 1$ es expuesto y $X = 0$ es no expuesto, los datos se expresan en la siguiente tabla.

	Caso	
Control	1	0
1	n_{11}	n_{01}
0	n_{10}	n_{00}

Así, n_{10} es el número de pares en los que el caso está expuesto y el respectivo control no. Para estos datos, la función de verosimilitud L de (9.19) es igual a

$$\begin{aligned} L(\beta) &= \left(\tfrac{1}{1+1}\right)^{n_{00}} \left(\tfrac{\exp(\beta)}{1+\exp(\beta)}\right)^{n_{10}} \left(\tfrac{1}{1+\exp(\beta)}\right)^{n_{01}} \left(\tfrac{\exp(\beta)}{\exp(\beta)+\exp(\beta)}\right)^{n_{11}} \\ &= \frac{\exp(\beta n_{10})}{[\exp(\beta) + \exp(\beta)]^{n_{10}+n_{01}}} \end{aligned} \tag{9.20}$$

De esta expresión, después de derivar respecto a β e igualar a 0, se obtiene un estimador puntual para $\exp(\beta)$ que corresponde a la razón de *odds RO*; esta es

$$\widehat{RO} = \exp(\widehat{\beta}) = \frac{n_{10}}{n_{01}} \tag{9.21}$$

la cual es idéntica a la estimación señalada en la ecuación (9.11).

Para verificar la influencia de la variable regresora X en la predicción de la variable dicotómica de interés Y, se contrasta la hipótesis:

$$H_0 : \ \beta = 0.$$

Esta hipótesis es equivalente a sostener que no existe relación entre la variable dependiente dicotómica Y y la variable regresora X.

Para verificar esta hipótesis, como la variable regresora X considerada es de tipo binario, se puede emplear la estadística de McNemar contenida en la expresión (9.15); para el caso de variables en escala de intervalo, se puede aplicar la estadística t o la estadística del rango signado de Wilcoxon.

Hosmer & Lemeshov (1989), en un estudio cuyo objetivo es identificar los factores de riesgo asociados al bajo peso al nacer (peso inferior a 2.500 gramos) en bebés (casos). Consideraron las variables sobre la madre: edad, peso a la última menstruación (en libras), hipertensión, tabaquismo e irritabilidad uterina. En las tres últimas variables 1 es sí y 0 es no. A cada uno de los casos (peso bajo al nacer) se le asignaron tres controles (peso normal al nacer), teniendo en cuenta la variable edad como criterio de emparejamiento. Con el propósito de ilustrar la estructura de los datos, en la tabla 9.8 se exhibe una parte de estos. Con relación a estos datos, suponga que se quiere indagar por la relación entre el bajo peso al nacer y el peso de la madre en el momento de su último periodo menstrual. Mediante el procedimiento PHREG[4] del SAS se obtienen los siguientes resultados:

Variable	Coeficiente	Error estándar	Estad. Z	valor p
Peso/madre	-0.0211	0.0112	-1.884	0.0593

De acuerdo con este $p-$valor (0.0593) se puede afirmar que hay un efecto

[4]La verosimilitud de esta regresión equivale a la del análisis de sobrevivencia estratificado.

Tabla 9.8: Bajo peso al nacer.

Conjunto emparejado	Caso	Peso/mad.	Hiperten.	Tabaq.	Irrit.-uter.
1	1	130	0	0	0
	0	112	0	0	0
	0	135	1	0	0
	0	270	0	0	0
2	1	110	0	0	0
	0	103	0	0	0
	0	113	0	0	0
	0	142	0	1	0
3	1	110	1	0	0
	0	100	1	0	0
	0	120	1	0	0
	0	229	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮

significativo entre el peso de la madre registrado y la ocurrencia del evento "nacimiento con bajo peso".

La razón de *odds* estimada, para un incremento de por ejemplo 10 libras de peso, de acuerdo con (9.21), es

$$\begin{aligned}\widehat{RO} &= \exp(\widehat{\beta}) \\ &= \exp(-0.0211) = 0.98,\end{aligned}$$

de manera que si una madre incrementa su peso en 10 libras, los *odds* a tener un bebe con bajo peso se reduce casi en un 2%.

La sintaxis SAS para efectuar estos cálculos se muestra en el siguiente recuadro.

```
DATA  BJO_PESO;
INPUT Conjunto Caso Peso_ma Hipert tabaq Irri_ute;
T_DUMMY=2-CASO;
CARDS;
    1        130        0          0       0
    0        112        0          0       0
    0        135        1          0       0
    0        270        0          0       0
```

```
    .          .          .          .          .
    .          .          .          .          .
    .          .          .          .          .
;
PROC  PHREG;
MODEL T_DUMMY*Caso(0)=Peso_ma/ TIES=DISCRETE;
STRATA=Conjunto;
```

donde `T_DUMMY` es una especie de "trampa" o "engaño" al procedimiento, pues este la exige así. La variable `Caso` indica que se trata de un caso o un control (1 o 0), y `Peso_ma` es la variable que registra el peso de la madre en el momento de la última menstruación. La variable `Conjunto` corresponde a la variable de estratificación (emparejamiento).

9.4.2 Regresión logística múltiple

En el ejemplo del bajo peso al nacer se consideran varios factores, los cuales estarían asociados a esta variable. En consecuencia, se debe ampliar el modelo propuesto anteriormente a uno en el que se pongan en juego varios factores (regresores) estrechamente relacionados con la variable dependiente. Este tipo de modelos involucra combinaciones lineales de variables regresoras. Como se muestra en el capítulo 5, estas variables pueden ser cuantitativas o categóricas.

El esquema *casos-controles* corresponde a un diseño en el que un conjunto de n_i casos se le asocian m_i controles (tabla 9.9). El desarrollo se hace en forma semejante al seguido para el caso de la regresión simple. La probabilidad condicional de observar una respuesta, dado que

Tabla 9.9: Emparejamiento $n_i : m_i$.

Conjunto	Caso	Control
	1	1
	2	2
i-ésimo	$\vdots$	$\vdots$
	$\vdots$	$\vdots$
	n_i	m_i

se tiene un conjunto con n_i casos, es

$$\frac{\exp\left(\sum_{j=1}^{n_i}(\beta^T x_i)\right)}{\sum_{R(n_i,m_i)}\exp\left(\sum_{j=1}^{n_i}(\beta^T x_i)\right)}, \tag{9.22}$$

La expresión $R(n_i, m_i)$ se refiere al número de formas en que se puede asignar controles a los casos, de manera que se tengan n_i casos y m_i controles. $R(n_i, m_i)$ corresponde a las posibles particiones de dos clases que se pueden efectuar con los $n_i + m_i$ individuos, de manera que una tenga n_i casos y la otra m_i controles. Además, $\beta^T = (\beta_1, \beta_2, \ldots, \beta_p)$ es el vector de parámetros. La función de verosimilitud completa es el producto de todos (N) los conjuntos emparejados (tabla 9.9):

$$L = \prod_{i=1}^{N} \frac{\exp\left(\sum_{j=1}^{n_i}(\beta^T x_i)\right)}{\sum_{R(n_i,m_i)}\exp\left(\sum_{j=1}^{n_i}(\beta^T x_i)\right)}. \tag{9.23}$$

De manera análoga al caso univariado, $\exp(\beta_i)$ representa:

(i) La razón de *odds* (RO) asociada a una exposición si X_i es binaria $(X_i = 1$ expuesto frente a $X_i = 0$ no expuesto).

(ii) La razón de *odds* (RO) debida a un incremento unitario si X_i es continua $(X_i = x + 1$ frente a $X_i = x)$.

También, una vez que $\widehat{\beta}_i$ y su error estándar han sido estimados, un intervalo del 95% de confiabilidad para la razón de *odds* anterior es dado por

$$\exp[\widehat{\beta}_i \mp 1.96\,\mathrm{ee}(\widehat{\beta}_i)]. \tag{9.24}$$

Una hipótesis que se puede verificar es la relacionada con la *contribución global* de las variables explicativas. Esta hipótesis se resume en la siguiente expresión:

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Esto significa que las p variables, consideradas en forma conjunta, no contribuyen a explicar la variación de la respuesta.

Se dispone de tres estadísticas con las que se verifica esta hipótesis, las cuales, bajo H_0, tienen distribución asintótica ji–cuadrado con p grados de libertad. Estas son:

1. La razón de verosimilitud

$$\chi^2_{RV} = 2[\ln L(\widehat{\beta}) - \ln L(\mathbf{0})]. \tag{9.25a}$$

2. La estadística de Wald

$$\chi^2_W = \widehat{\beta}^T \left(V(\widehat{\beta}) \right)^{-1} \widehat{\beta}, \tag{9.25b}$$

donde $\widehat{\beta}^T$ es el traspuesto del vector $\widehat{\beta}$ y $V(\widehat{\beta})$ es la matriz de covarianzas del vector $\widehat{\beta}$.

3. La prueba de puntaje

$$\chi^2_S = \left[\frac{\partial \ln L(\mathbf{0})}{\partial \beta} \right]^T \left[\frac{\partial^2 \ln L(\mathbf{0})}{\partial \beta^2} \right]^{-1} \left[\frac{\partial \ln L(\mathbf{0})}{\partial \beta} \right] \tag{9.25c}$$

Con relación al ejemplo sobre bajo peso al nacer, el modelo de regresión logística condicional con las cuatro variables: edad, peso a la última menstruación (en libras), hipertensión, tabaquismo, e irritabilidad uterina, la tabla 9.10 contiene las respectivas estimaciones de los parámetros.

La hipótesis que da cuenta de la no contribución de estas cuatro variables al evento "peso bajo al nacer" es

$$H_0 : \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.$$

La respectivas estadísticas de verificación junto con los $p-$valores se consignan en el siguiente cuadro.

Estadística	G.L.	$p-$valor
$\chi^2_{RV} = 9.530$	4	0.0491
$\chi^2_W = 6.001$	4	0.1991
$\chi^2_S = 8.491$	4	0.0752

Se puede leer, desde estos resultados, que una o más de estas variables explicativas se asocian de manera moderadamente significativa, excepto la estadística de Wald, con la respuesta de interés.

Se quiere verificar ahora acerca de la afectación parcial de una variable particular; también se desea verificar si la adición de una variable al modelo es significativa, dado que las demás han sido previamente incluidas en el modelo. La hipótesis a verificar tiene la forma

$$H_0 : \ \beta_{i|j\neq i} = 0, \ \text{para } i, j = 1, 2, \ldots, p. \tag{9.26}$$

Otra lectura de esta hipótesis es: "La variable X_i no tiene influencia adicional sobre la respuesta dado que otras variables ya han sido incluidas en el modelo".

La estadística con la cual se verifica la hipótesis (9.26) es

$$Z_i = \frac{\widehat{\beta}_i}{\text{ee}(\widehat{\beta}_i)}, \tag{9.27}$$

donde $\widehat{\beta}_i$ es el estimador del coeficiente y $\text{ee}(\widehat{\beta}_i)$ el respectivo error estándar parcial. El paquete SAS, a través del procedimiento LOGISTIC o PHREG, suministra estos cálculos. En la tabla 9.10 se muestran las estimaciones, la estadística Z_i anterior y el respectivo p-valor. Otro

Tabla 9.10: Estimación para datos peso bajo al nacer.

Variable	*Estimación*	*Error estánd.*	Estad. Z_i	p-valor
Peso madre	-0.0191	0.0114	-1.673	0.0942
Tabaquismo	-0.0191	0.0114	-0.103	0.9182
Hipertensión	-0.0191	0.0114	0.528	0.5975
Irrit. uterina	-0.0191	0.0114	1.784	0.0745

tipo de verificación que se puede hacer está relacionada si un subgrupo de k variables ($k < p$) contribuye sobre la predicción a condición de que el modelo contenga previamente las otras variables. Como se observa, es una extensión del caso anterior. La hipótesis a verificar se expresa como sigue:

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_k = 0. \tag{9.28}$$

Para verificar esta hipótesis se obtiene la estadística de razón de verosimilitud con k grados de libertad; esta es:

$$\chi^2_{RV} = 2[\ln L_{H_1}(\widehat{\beta}) - \ln L_{H_0}(\widehat{\beta}] \tag{9.29}$$

donde $L_{H_1}(\widehat{\beta})$ representa la verosimilitud del modelo con todas las X y $L_{H_0}(\widehat{\beta})$ es la verosimilitud del modelo sin las variables X de interés.

Para un estudio puede resultar de importancia observar la contribución conjunta de subgrupos de variables morfológicas, fisiológicas, demográficas, económicas, entre otras. Una aplicación adicional del contraste anterior es la consideración de potencias o productos entre las variables (a la manera de interacciones).

Para los datos sobre bajo peso al nacer[5] (por última vez), considere las siguientes interacciones:

- (Peso madre)*(Tabaquismo).

- (Peso madre)*(Hipertensión).

- (Peso madre)*(Irritabilidad uterina)

Con estas interacciones se trata de observar si alguna de las demás variables modifica el efecto del peso de la madre en la respuesta (tener bebés con bajo peso).

De acuerdo con la estadística (9.29), se debe calcular la verosimilitud en las cuatro variables originales para restarla de la verosimilitud en todas las siete variables. Así:

a) Para las cuatro variables $\ln L = -16.030$

b) Para todas las variables, cuatro originales y tres interacciones adicionales, $\ln L = -14.199$

El valor de la estadística (9.29) es

$$\begin{aligned}\chi^2_{RV} &= 2[\ln L(\widehat{\beta};\ \text{en las siete variables}) - \ln L(\widehat{\beta};\ \text{con las cuatro variables})] \\ &= 3.62\end{aligned}$$

la cual, para una ji–cuadrado con 3 grados de libertad, advierte que estas interacciones son débiles (además el valor p es mayor que 0.10).

Además de las hipótesis anteriores con el propósito de encontrar el modelo que "mejor" se ajuste a los datos, se puede optar por la selección de variables a través de los procedimientos: regresión hacia adelante (*forward*), regresión hacia atrás (*backward*) y regresión paso a paso (*stepwise*). El paquete SAS suministra las respectivas herramientas computacionales en los procedimientos LOGISTIC y PHREG.

[5] Le (1998, Cap. 5).

9.5 Procesamiento de datos con R

Medidas de concordancia o acuerdo

En esta sección se explica cómo calcular la estadística κ de acuerdo con la ecuación (9.4). El código en R es el siguiente:

```
# introducción de los datos
x<-c(38,33,10,3, 5,11,14,7, 0,3,5,3, 1,0,6,10)
# conversión a un objeto de tabla (table)
tabla9.3<-as.table(matrix(x,nrow=4))
# Nombres de las categorías
dimnames(tabla9.3)<-list(
c("Seguro","Muy probable","Dudoso","Improbable"),
c("Seguro","Muy probable","Dudoso","Improbable") )
# en forma de proporciones
tabla9.3p<-prop.table(tabla9.3)
# marginal de columna
pi.<-margin.table(tabla9.3p,1)
# marginal de fila
p.i<-margin.table(tabla9.3p,2)
# proporciones en la diagonal de la tabla
pii<-diag(prop.table(tabla9.3p))
# estadística kappa
(sum(pii)-sum(pi.*p.i) )/( 1-sum(pi.*p.i))
```

Regresión logística condicional

Para ajustar modelos de regresión logística condicional con R, se cuenta con la función `clogit()` de la librería `survival`. Para detalles de su uso, el lector puede consultar la ayuda de la librería. En el capítulo 16 de Virasakdi (2004), se pueden consultar otros ejemplos.

9.6 Ejercicios

1. La información que se transcribe en la tabla 9.11 forma parte de un estudio acerca de la toma de anticonceptivos orales entre 183 mu-

jeres con cáncer endometria y 183 sin este. Su objetivo es analizar la posible influencia de estos medicamentos en el desarrollo de la citada neoplasia.

a) ¿Cuál es el riesgo de presentar la neoplasia entre las que toman anticonceptivos orales frente a las que no los toman?

b) Estime la razón de *odds* mediante el procedimiento de Cochran-Mantel-Haenszel.

Tabla 9.11: Influencia de los anticonceptivos orales sobre el cáncer endometrial.

	Controles		
Casos	Toma ACO	No toma ACO	Total
Toma ACO	12	43	55
No toma ACO	7	121	128
	19	164	183

Fuente: Case-control studies, Schlesselman 1982.

2. Los datos del archivo `bbdm13.dat`[6] corresponden a un estudio sobre enfermedad benigna de la mama con un emparejamiento 1:3. Para detalles del experimento el lector puede consultar (Hosmer & Lemeshov 1989). Encuentre el mejor modelo de regresión logística para un emparejamiento 1:1 usando el primero de los tres controles.

[6]Tomados de `http://www.umass.edu/statdata/statdata/data/bbdm13.dat`

Apéndice A

Tablas

Tabla A.1: Distribución normal acumulada: $\int_z^{-\infty} \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}dt$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

Tabla A.2: Percentiles de la distribución ji–cuadrado.

	P				
ν	0.005	0.01	0.025	0.05	0.1
1	7.879	6.635	5.024	3.841	2.706
2	10.597	9.210	7.378	5.991	4.605
3	12.838	11.345	9.348	7.815	6.251
4	14.860	13.277	11.143	9.488	7.779
5	16.750	15.086	12.833	11.070	9.236
6	18.548	16.812	14.449	12.592	10.645
7	20.278	18.475	16.013	14.067	12.017
8	21.955	20.090	17.535	15.507	13.362
9	23.589	21.666	19.023	16.919	14.684
10	25.188	23.209	20.483	18.307	15.987
11	26.757	24.725	21.920	19.675	17.275
12	28.300	26.217	23.337	21.026	18.549
13	29.819	27.688	24.736	22.362	19.812
14	31.319	29.141	26.119	23.685	21.064
15	32.801	30.578	27.488	24.996	22.307
16	34.267	32.000	28.845	26.296	23.542
17	35.718	33.409	30.191	27.587	24.769
18	37.156	34.805	31.526	28.869	25.989
19	38.582	36.191	32.852	30.144	27.204
20	39.997	37.566	34.170	31.410	28.412
21	41.401	38.932	35.479	32.671	29.615
22	42.796	40.289	36.781	33.924	30.813
23	44.181	41.638	38.076	35.172	32.007
24	45.559	42.980	39.364	36.415	33.196
25	46.928	44.314	40.646	37.652	34.382
26	48.290	45.642	41.923	38.885	35.563
27	49.645	46.963	43.195	40.113	36.741
28	50.993	48.278	44.461	41.337	37.916
29	52.336	49.588	45.722	42.557	39.087
30	53.672	50.892	46.979	43.773	40.256

Apéndice B

Procedimientos básicos con R

`R` es un sistema para realizar calculos y gráficos estadísticos, cuenta con un poderoso lenguaje de programación orientado a objetos, un entorno para la ejecución de comandos, un depurador de código de programación, el acceso a determinadas funciones del sistema y la capacidad de ejecutar programas almacenados en archivos (*scripts*). `R` fue escrito por Ross Ihaka y Robert Gentleman en el Departamento de Estadística de la Universidad de Auckland, en Auckland, Nueva Zelanda. Además, un gran grupo de personas ha contribuido enviando código de programación e informes de error. `R` es *software* libre, se puede descargar desde `www.r-project.org/`, donde se consigue en versiones para los sistemas operativos Windows®, Linux® y Macintosh®.

En este apéndice se explica cómo hacer algunas tareas usuales que se requieren en el análisis de datos, efectuar cálculos de probabilidades y cuantiles, leer datos externos, ordenarlos y transformarlos.

B.1 Cálculo de probabilidades y cuantiles

Cuando se cuenta con un paquete para cálculo estadístico como `R`, las tablas del apéndice A no son necesarias. En esta sección y la que sigue,

se presenta la forma de calcular probabilidades y percentiles a partir de las distribuciones binomial, de Poisson, normal y ji–cuadrado.

B.1.1 Distribución binomial

Antes de pasar a la distribución binomial, se ilustra cómo ralizar cálculo de combinaciones; la función es `choose()`. Veamos cómo se usa en tres ejemplos:

El valor de	Se obtiene mediate
$\binom{10}{3}$	`choose(10,3)`
$\binom{15}{10}$	`choose(15,10)`
$\binom{100}{4}$	`choose(100,4)`

Cálculo de probabilidades a partir de la distribución binomial

Ejemplo: si X es una variable aleatoria con distribución binomial de parámetros $n = 18$ y $p = 0.1$, calcule $P(X = 2)$.

```
dbinom(x=2,size=18,prob=0.1)
```

Ejemplo: si X es una variable aleatoria con distribución binomial de parámetros $n = 18$ y $p = 0.1$, calcule $P(X \geq 4) = 1 - P(X \leq 3)$.

```
1-pbinom(3,size=18,p=0.1)
```

Esta probabilidad es equivalente a $P(X > 3)$, la cual se calcula mediante

```
pbinom(3,18,0.1,lower.tail=FALSE)
```

también es equivalente a $\sum_{x=4}^{18} P(X = x)$, que se calcula mediante

```
sum( dbinom(4:18,18,0.1) )
```

Ejemplo: si X es una variable aleatoria con distribución binomial de parámetros $n = 18$ y $p = 0.1$, calcule $P(3 \leq X < 7)$. Esta probabilidad es igual a $= F(6) - F(2)$, se procede como sigue

```
pbinom(6,18,0.1)-pbinom(2,18,0.1)
```

lo anterior es equivalente a

```
sum( dbinom(3:6,18,0.1) )
```

Cálculo de cuantiles a partir de la distribución binomial

Si X es una variable aleatoria binomial de parámetros n y p, la función `qbinom()` regresa el valor más pequeño x tal que $P(X \leq x) \geq \alpha$. Veamos un ejemplo:

```
qbinom(0.99,18,0.1)
[1] 5
```

significa que $P(X \leq 5) \geq 0.99$ y 5 es el valor más pequeño que cumple esa desigualdad.

Para la simulación de muestras aleatorias de la distribución binomial se usa la función `rbinom()`.

B.1.2 Distribución de Poisson

Las funciones para la distribución de Poisson son: `dpois()` para probabilidad puntual, `ppois()` para probabilidad acumulada, `qpios()`, para cuantiles y `rpois()` para generar muestras aleatorias.

Cálculo de probabilidades a partir de la distribución de Poisson

Ejemplo: si X es una variable aleatoria con distribución de Poisson con media 2.3, calcule $P(X = 2)$.

```
dpois(x=2,lambda=2.3)
```

Ejemplo: si X es una variable aleatoria con distribución de Poisson con media $E(X) = 4.6$, calcule $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 0)$.

```
1-ppois(0,4.6)
```

Cálculo de cuantiles a partir de la distribución de Poisson

Similar al caso de la distribución binomial. Si X es una variable aleatoria Poisson de parámetro λ, la función `qpois()` regresa el valor más pequeño x tal que $P(X \leq x) \geq \alpha$. Por ejemplo:

```
qpois(0.99,2.3)
[1] 6
```

significa que $P(X \leq 6) \geq 0.99$ y 6 es el valor más pequeño que cumple esa desigualdad.

El lector habrá notado el patrón usado por `R`. Para el cálculo de probabilidad puntual se usa `d` seguido del nombre de la función (`dbinom`, `dpois`); para probabilidad acumuladoa se usa `p` seguido del nombre de la función; para cuantiles se usa `q`; y para la generación de datos aleatorios se usa `r`. De esa forma es fácil deducir cómo calcular estos valores con las otras distribuciones, por ejemplo `dgeom()`, `pgeom()`, `qgeom()` y `rgeom()` para la distribución geométrica; `dhyper()`, `phyper()`, `qhyper()` y `rhyper()` para la distribución hipergeométrica.

B.1.3 Distribuciones normal y ji–cuadrado

La sintaxis de `R` para el cálculo de probabilidades, cuantiles y la generación de números aleatorios, en el caso de las variables aleatorias continuas, es similar al caso discreto. En este el prefijo `d`, en lugar de calcular probabilidad puntual (que no tiene sentido en el caso continuo), evalúa la función de densidad correspondiente. En este libro se usa frecuentemente las distribuciones normal y ji–cuadrado; por eso se estudian con detalle en las dos secciones siguientes.

Distribución normal

Para la distribución normal se cuenta con las funciones `dnorm()`, `pnorm()`, `qnorm()` y `rnorm()`. Veamos cómo usarlas por medio de ejemplos.

Si Z es una variable aleatoria normal estándar, calcule $P(Z \leq 1.65)$.

```
pnorm(1.65)
[1] 0.9505285
```

Note que por defecto la función regresa la probabilidad acumulada desde $-\infty$ hasta el valor dado; además, toma media cero y varianza 1 (normal estándar). En general, si la variable aleatoria no es normal estándar, hay que especificar la media y la desviación estándar. Por ejemplo, si X es una variable aleatoria con media 10 y varianza 4, calcule $P(X < 14)$.

```
pnorm(14,mean=10,sd=2)
[1] 0.9772499
```

Si se desea $P(X > 14)$ se puede hacer de dos formas: $1 - P(X < 14)$ o mediante el comando

```
pnorm(14,mean=10,sd=2,lower.tail=FALSE)
[1] 0.02275013
```

Obtener un valor z tal que $P(Z < z) = 0.99$

```
qnorm(0.99)
[1] 2.326348
```

Distribución ji–cuadrado

Para la distribución ji–cuadrado se cuenta con las funciones `dchisq()`, `pchisq()`, `qchisq()` y `rchisq()`. Veamos cómo usarlas por medio de ejemplos.

Si X^2 es una variable aleatoria ji–cuadrado con un grado de libertad, calcule $P(X^2 > 3.84) = 1 - P(X^2 \leq 3.84)$.

```
1-pchisq(3.84,df=1)
[1] 0.05004352
```

Si X^2 es una variable aleatoria ji–cuadrado con dos grados de libertad, halle el valor x tal que $P(X^2 > x) = 0.05$. Lo que se pide es equivalente a encontrar x tal que $P(X^2 < x) = 0.95$ y se obtiene mediante el comando

```
qchisq(0.95,df=2)
[1] 5.991465
```

Para las demás distribuciones continuas de probabilidad la sintaxis es similar. Para la uniforme, `dunif()`, `punif()`, `qunif()` y `runif()`; para la distribución t de student, `dt()`, `pt()`, `qt()` y `rt()`; para la distribución F, `df()`, `pf()`, `qf()` y `rf()`, el usuario solo debe tener en cuenta los parámetros que definen las distribuciones.

B.2 Lectura de datos externos

Una fuente de error frecuente para el que está iniciándose en R es el olvido de fijar el *directorio de trabajo*, que es el directorio donde R buscará, por defecto, los archivos a leer. Si el archivo de datos que se pretende almacenar en la memoria no se encuentra en el directorio de trabajo, hay que especificar la ruta exacta del archivo, para que la lectura sea exitosa. Para no tener que digitar la ruta completa de ubicación de los archivos, se recomienda crear, para cada proyecto, un directorio donde se colocarán todos los archivos de datos y código asociados.

B.2.1 El directorio de trabajo

La función que permite conocer el directorio de trabajo actual es `getwd()`.[1] Al llamar a esta función, `R` regresa la ruta del directorio de trabajo, como se muestra a continuación (la salida cambiará de acuerdo con la configuración del sistema).

```
> getwd()
[1] "/home/mario"
```

Para cambiar el directorio de trabajo se usa la función `setwd()`. Por ejemplo, si se desea fijar la carpeta `/home/mario/proyecto` como direc-

[1] *get working directory.*

torio de trabajo, se procede como sigue:[2]

```
> setwd("/home/mario/proyecto")
```

B.2.2 Lectura de datos desde un archivo de texto

Cuando se tienen los datos en un archivo de texto plano, se cuenta con la función `read.table()`. Este tipo de archivo se reconoce por la extensión, que usualmente es `.txt` o `.dat`. Suponga que se desea leer el archivo[3]

```
height weight
  58    115
  59    117
   .     .
   .     .
   .     .
  71    159
  72    164
```

el cual se tiene guardado en el directorio de trabajo con el nombre `archivo.txt`. El comando es

```
datos<-read.table("archivo.txt",header=TRUE)
```

De esta forma, se crea el objeto `datos`, que contiene la información almacenada en `archivo.txt`. La opción `header=TRUE` es indispensable porque el archivo tiene, en su primera fila, los nombres de las columnas.

B.2.3 Lectura de datos desde un archivo CSV

Un archivo CSV es un archivo de texto en el cual los datos están separados por comas. La ventaja que tiene trabajar con este formato es que

[2] En la versión para Windows, además del comando, el usuario cuenta con el menú archivo > fijar directorio de trabajo.

[3] *Data frame* `women` del paquete base.

todas las hojas de cálculo son capaces de crearlos y leerlos. Es frecuente que el dueño de la información tenga sus datos organizados en una hoda de cálculo de Open Office (.odt) o Excel (.xls). Para importar dichos datos a la memoria de R, podemos crear un archivo con formato CSV y leerlo con la función `read.csv()` o con `read.csv2()`. Suponga que deseamos leer los siguientes datos, guardados en el directorio de trabajo con el nombre `ejemplocsv.csv`.

```
tiempo, momento, resp
 1, 0 ,95
 2, 0 ,93
 .  .  .
 .  .  .
 .  .  .
19, 1 ,7
20, 1 ,3
```

Note que las columnas están separadas por comas. La lectura de estos datos se hace mediante el comando

```
read.csv("ejemplocsv.csv")
```

Si en lugar de la coma tenemos punto y coma separando las columnas, la función apropiada es `read.csv2()`. Por ejemplo, si tenemos los mismos datos anteriores, pero separados por `;` en un archivo ubicado en el directorio de trabajo, con nombre `ejemplocsv2.csv`

```
tiempo ; momento ; resp
 1; 0 ;95
 2; 0 ;93
 .  .  .
 .  .  .
 .  .  .
19; 1 ;7
20; 1 ;3
```

la lectura se hace mediante el comando

```
read.csv2("ejemplocsv2.csv")
```

La diferencia entre `read.csv()` y `read.csv2()` radica esencialmente en el caracter usado como separador de campos, como se ilustra en el ejemplo anterior, y en el usado para el separador decimal; la primera usa `dec=","` mientras que la segunda usa por defecto `dec="."`. En realidad, el segundo archivo de datos se puede leer con el comando

```
read.csv("ejemplocsv2.csv",sep = ";")
```

B.2.4 Lectura de datos desde un archivo de Excel

Existen varias formas de leer un archivo en formato de Excel (`.xls`). Aquí se ilustra cómo hacerlo con las librerías `RODBC` y `gdata`. Suponga que tenemos el archivo `ejemplo.xls` con varias hojas y que los datos de interés se encuentran en la hoja 3. La lectura, mediante `RODBC`, se hace con el siguiente código:

```
library(RODBC) #debe estar instalada previamente
archivo<-odbcConnectExcel("ejemplo.xls")
tabla<-sqlFetch(archivo,"Hoja3")
close(archivo)
```

de esa forma se crea el objeto `tabla` que contiene los datos de la hoja 3. La orden `close(archivo)` es para cerrar la conexión con el archivo. Mientras la conexión esté activa no es posible hacer cambios en el archivo `ejemplo.xls`. La lectura de la misma hoja del archivo, pero usando la librería `gdata` se indica a continuación

```
library(gdata) #debe estar instalada previamente
tabla<-read.xls("ejemplo.xls",sheet=3)
```

la función `read.xls()` trabaja convirtiendo, de manera temporal, el archio de excel a un archivo csv.

B.3 Selección y transformación de datos

Una vez cargado un conjunto de datos en la memoria de `R`, los pasos siguientes son: selección de un subconjunto de ellos, creación de nuevas

variables a partir de las antiguas y obtención de algunas estadísticas básicas bien sea en forma global o de acuerdo con algún criterio. En esta sección se ilustran, a través de ejemplos, los comandos básicos para realizar estas tareas. Para los ejemplos se usará la base de datos `survey` de la librería `MASS`, que se carga a la memoria con los siguientes comandos

```
library(MASS)
data(survey)
```

Para imprimir las primeras y las últimas 6 filas de los datos, se usan los siguientes comandos:

```
head(survey)
tail(survey)
```

Las estadísticas descriptivas básicas se obtienen con

```
summary(survey)
```

Cuando la variable es de caracteres o un factor, el resumen consta de una lista de los niveles del factor acompañada de la frecuencia de cada nivel. Se observa que algunas variables tienen datos faltantes. Para eliminar del archivo las filas con datos faltantes, se procede como sigue:

```
datos<-na.omit(survey)
summary(datos)
```

De esta forma creamos el objeto datos que no contienen datos faltantes (`NA`).

B.3.1 Creación de nuevas variables

Suponga que se requiere crear una nueva columna, a partir de las existentes en `datos`, mediante una transformación. Por ejemplo, para crear una columna que contenga el logaritmo natural de `Height`, procedemos así:

```
datos$lHeight<-log(datos$Height)
```

Otras funciones de uso frecuente son `sqrt()` (raíz cuadrada), `exp()` (exponencial), `abs()` (valor absoluto) y `sin()` (seno).

B.3.2 Selección de subconjuntos de un marco de datos

A continuación se explica cómo usar el sistema de indexación de `R` para la selección de filas y columnas de un marco de datos.

Selección de columnas

El comando

```
datos[,1:4]
```

permite seleccionar las primeras 4 columnas del archivo datos. Note el uso de la coma, la cual separa filas de columnas (`[filas,columnas]`). En este caso no hay ningún criterio para las filas; por tanto, se seleccionan todas. Si se quiere seleccionar las columnas numéricas, que están en las posiciones 2, 3, 6, 10 y 12, procedemos así:

```
datos[,c(2,3,6,10,12)]
```

Las columnas también se pueden seleccionar por sus nombres, como se indica en el siguiente código:

```
datos[,c("Wr.Hnd","NW.Hnd","Pulse","Height","Age")]
```

Para seleccionar las columnas que no son numéricas, dado que se conoce cuáles son las numéricas, se usa el siguiente código (note el uso del signo menos):

```
datos[,-c(2,3,6,10,12)]
```

La siguiente forma es interesante, porque la selección de las columnas es automática: si se quiere seleccionar todas las variables numéricas (sin hacer una lista exhaustiva de ellas o de sus posiciones), se procede así:

```
numericas<-sapply(datos, is.numeric)
datos[,numericas]
```

La función `sapply` entrega `TRUE` donde hay una columna numérica y `FALSE` en caso contrario. El objeto `numericas` es lo que se conoce como un vector de tipo lógico. Mediante el comando `datos[,numericas]`, `R` selecciona las columnas cuya posición corresponda a un `TRUE` en el vector lógico `numericas`. Para seleccionar las columnas que no sean numéricas, se usa la orden

```
datos[,!numericas]
```

El signo de exclamación se usa en `R` para la negación lógica.

Selección de filas

El comando

```
datos[1:7,]
```

selecciona las 7 primeras filas del archivo datos. Si se desea seleccionar solo las 7 primeras filas de las columnas numéricas, se procede así:

```
datos[1:7,numericas]
```

El comando

```
datos[c(3,5,1,4,9), ]
```

selecciona las filas 3, 5, 1, 4 y 9, en ese orden. Una situación que se presenta a menudo es la selección de individuos (filas) de acuerdo con un criterio; por ejemplo, seleccionar a los hombres o seleccionar a las mujeres que fuman. A continuación se muestran ejemplos concretos.

```
datos[datos$Sex=="Male", ]
```

selecciona todos los individuos hombres. El siguiente comando selecciona las mujeres que nunca fuman (note el uso del operador lógico `&`):

```
datos[Sex=="Female" & Smoke=="Never",]
```

Ordenar una base de datos

A veces es necesario ordenar una base de datos por algún criterio. Aquí se explica cómo hacerlo

```
datos[order(datos$Age),]
```

ordena el archivo datos por la columna edad. En este caso lo hace en forma ascendente; si se quiere en orden descendente, se usa (note el signo menos):

```
datos[order(-datos$Age),]
```

Si se requiere ordenar por más de un criterio, digamos edad y altura, se procede como se indica a continuación:

```
datos[order(datos$Age,datos$Height),]
```

B.3.3 Cálculos por niveles de un factor

Si se quiere calcular alguna función (por ejemplo la media), para cada nivel de un factor (por ejemplo el sexo), en `R` se cuenta con varias funciones. Veremos cómo usar la función `by()`.

```
by(datos$NW.Hnd,datos$Sex,mean)
```

calcula la media de la variable `NW.Hnd` para cada nivel de sexo. La orden

```
lapply(datos[,numericas],var)
```

calcula la varianza de cada una de las variables numéricas de datos. La línea de comando

```
sapply(datos[,numericas], var )
```

hace exactamente lo mismo que la anterior, la única diferencia es la forma como `R` organiza la respuesta; en el primer caso, el objeto regresado por la función es una lista, mientras que en el segundo es un vector de tipo

numérico. La siguiente línea de código regresa los mismos valores; en este caso, el número 2 indica que la varianza se va a calcular sobre la segunda dimensión del objeto, es decir, las columnas. Si se usa 1 en lugar de 2, calcula la media para cada fila. La ventaja de `apply` es que esta función puede emplearse en cálculos con arreglos multidimensionales (`arrays`).

```
apply(datos[,numericas],2,var)
```

Si se quiere una tabla de contingencia cruzando las variables sexo y ejercicio, se puede usar la orden

```
tapply(datos$Sex,datos[,c("Sex","Exer")],length)
```

Lo anterior es equivalente a

```
table(datos$Sex,datos$Exer)
```

solo que es ventajoso usar `tapply()` porque en el lugar de `length` se puede usar otra función, como `sum` o `var`. Por ejemplo, la edad promedio para cada nivel de sexo y ejercicio se obtiene mediante

```
tapply(datos$Age,datos[,c("Sex","Exer")],mean)
```

La función `agregate()` realiza el mismo trabajo de la función `by()`, solo que la salida de los resultados es un objeto de clase `data.frame`, lo cual es más elegante en cuanto a la presentación.

```
aggregate(datos[,numericas], list(datos$Sex),mean)
```

Con el comando anterior se obtiene la media de las variables numéricas para cada nivel de sexo. El comando siguiente obtiene la mediana de las variables numéricas para cada combinación de niveles de los factores `Sex` y `W.Hnd`.

```
aggregate(datos[,numericas],list(datos$Sex,datos$W.Hnd),
          median)
```

En los dos ejemplos anteriores las funciones `mean` y `median` se pueden reemplazar por cualquier otra que regrese un valor simple, por ejemplo `max`, `var`, `length`. Ejemplos de funciones que, en este caso, no se pueden usar con `aggregate`: `cov` y `cor`, ya que el objeto que se entrega es de la clase `data.frame` y el resultado sería una matriz, esa es una desventaja en relación con la función `by`.

Bibliografía

Agresti, A. (1996), *An introduction to categorical data analysis*, John Wiley, New York.

Agresti, A. (2000), *Categorical data analysis*, John Wiley, New York.

Andersen, E. B. (1997), *Introduction to the statistical analysis of categorical data*, Springer, New York.

Ato, M. & López, J. J. (1996), *Análisis estadístico para datos categóricos*, Síntesis S. A., Madrid.

Benzecri, J. P. (1973), *L'analyse des donnés, Tomo 1: La taxinomie, Tomo 2: L'Analyse des correspondances*, Dunod, Paris.

Bishop, Y. V., Fienberg, S. E. & Holland, P. W. (1975), *Discrete multivariate analysis*, MA: MIT Press, Cambridge.

Carlin, B. P. & Louis, T. A. (1998), *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall/CRC, London.

Christensen, R. (1990), *Log-linear models*, Springer-Verlag, New York.

Conover, W. J. (1990), *Estadística no paramétrica*, McGraw-Hill, New York.

Dobson, A. J. (2002), *An introduction to generalizaed linear models*, Chapman & Hall, New York.

Díaz, L. G. (2007), *Estadística multivariada: inferencia y métodos*, Universidad Nacional de Colombia, Bogotá.

Escofier, B. et Pages, J. (1990), *Analyses factorielles simples et multiples*, Dunod, Paris.

Everitt, B. S. (1994), *The analysis of contingency tables*, Chapman and Hall, London.

Gibbons D., J. (1971), *Nonparametric statistical inference*, MacGraw-Hill Kogakusha, Ltda.

Gil, S. I. & Zárate, G. P. (1984), *Métodos Estadísticos*, Trillas.

Hettmansperger, T. P. (1984), *Statistical inference based on ranks*, John Wiley and Sons.

Hosmer, D. W. & Lemeshov, S. (1989), *Applied logistic regression*, John Wiley and Sons, New York.

Juez, P. & Díez, J. (1997), *Probabilidad y estadística en medicina*, Díaz de Santos S. A., Madrid.

Kendall, M. & Stuart, A. (1979), *The advanced theory of statistics*, Vol. 2, Charles Griffin and Company, London.

Le, C. T. (1998), *Applied categorical data analysis*, John Wiley and Sons, New York.

Lebart, L., M. A. F. J. P. (1985), *Tratamiento estadístico de datos*, Marcombo-Boixareu Editores, Barcelona.

Morrison, D. F. (1990), *Multivariate statistical methods*, McGraw-Hill Book Company, New York.

Paulino, C. D. e Singer, J. M. (2006), *Analálise de dados categorizados*, Editora Edgar Blucher, Sao Paulo.

Peña, D. (1998), *Estadística modelos y métodos. 1. Fundamentos*, Alianza Universitaria Textos, Madrid.

Rencher, A. C. (1995), *Methods of multivariate analysis*, John Wiley and Sons, New York.

Rencher, A. C. (1998), *Multivariate statistical inference and applications*, John Wiley and Sons, New York.

Silva, L. C. (1995), *Excursión a la regresión logística en ciencias de la salud*, Díaz de Santos S. A., Madrid.

Stokes, M. E., Davis, C. S. & Koch, G. G. (1997), *Categorical data analysis using the SAS System*, As Institute Inc., New York.

Team, R. D. C. (2007), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Virasakdi, C. (2004), *Analysis of epidemiological data using R and Epicalc*, Prince of Songkla University., Thailand.

Índice temático

Análisis estadístico de datos categóricos
se terminó de imprimir en Editorial UN,
en septiembre de 2009.
Bogotá, D.C., Colombia.