EJEMPLOS SOBRE EL ANÁLISIS DE DATOS CATEGÓRICOS

MARÍA ELENA ROJAS GARCÍA

TESIS

Presentada Como Requisito Parcial
Para Obtener el Grado de

Maestro en Ciencias en
Estadística Experimental

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

PROGRAMA DE GRADUADOS

Buenavista, Saltillo, Coahuila, México.
Junio de 2007

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

SUBDIRECCIÓN DE POSTGRADO

EJEMPLOS SOBRE EL ANÁLISIS DE DATOS CATEGÓRICOS

TESIS

POR:

MARÍA ELENA ROJAS GARCÍA

Elaborada bajo la supervisión del comité particular de asesoría y aprobada como requisito parcial para optar al grado de

MAESTRO EN CIENCIAS EN
ESTADÍSTICA EXPERIMENTAL

COMITÉ PARTICULAR

Asesor Principal: _____
Dr. Rolando Cavazos Cadena

Asesor: _____
M. C. Luis Rodríguez Gutiérrez

Asesor: _____
M. C. Félix de Jesús Sánchez Pérez

_____
Dr. Jerónimo Landeros Flores
Subdirector de Postgrado

Buenavista, Saltillo, Coahuila. Junio de 2007

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

SUBDIRECCIÓN DE POSTGRADO

COMPENDIO

EJEMPLOS SOBRE EL ANÁLISIS DE DATOS CATEGÓRICOS

TESIS

POR:

MARÍA ELENA ROJAS GARCÍA

MAESTRÍA EN CIENCIAS EN
ESTADÍSTICA EXPERIMENTAL

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

Buenavista, Saltillo, Coahuila. Junio de 2007

Dr. Rolando Cavazos Cadena. Asesor

Palabras Clave: Datos categóricos, Estadístico suficiente, Verosimilitud máxima, Estimación y prueba sobre proporciones, Razón de oportunidades, Tasa de riesgo. Prueba de Independencia.

El objetivo de este trabajo es ilustrar la aplicación de ideas estadísticas básicas al análisis de datos categóricos mediante ejemplos detallados. La principal contribución es la formulación precisa de la idea de catagoría y la discusión detallada del estadístico suficiente en el análisis de este tipo de datos.

ABSTRACT

EXAMPLES ON CATEGORICAL DATA ANALYSIS

BY

MARÍA ELENA ROJAS GARCÍA

MAESTRÍA EN CIENCIAS EN
ESTADÍSTICA EXPERIMENTAL

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

Buenavista, Saltillo, Coahuila. Junio de 2007

Dr. Rolando Cavazos Cadena. Advisor

Key Words: Categorical data, Sufficient statistic, Maximum likelihood, Estimates and test on proportions, Odds ratio, Risk rate, Test of independence.

The objective of this work is to present detailed applications of basic statistical ideas to the analysis of categorical data. The main contribution is the formulation of a precise notion of data category, and the detailed discussion of the sufficient statistic in the analysis of categorical variables.

Índice de Contenido

1. Suficiencia y Método de Verosimilitud Máxima en el Análisis de Datos Categóricos

1.1 Categorías de datos 1

1.2 El Estadístico Suficiente 3

1.3 Método de Verosimilitud Máxima 6

1.4 Prueba de Razón de Verosimilitud 7

1.5 Prueba Ji-Cuadrada 10

2. Ejemplos Sobre las Ideas Básicas

2.1 Variables de Respuesta y Explicatorias 11

2.2 Ejemplos sobre las Distribuciones Binomial y de Poisson 13

2.3 Prueba de Hipótesis para un Parámetro Binomial 18

2.4 Ejemplos sobre la Función de Verosimilitud 23

2.5 Intervalo de Confianza Alternativo en el Caso Binomial 27

3. Medidas de Asociación

3.1 Clasificación Doble 29

3.2 Distribución Condicional 31

3.3 Medidas de Asociación 32

3.4 Prueba de Independencia 33

4. Ejemplos del Análisis de Tablas de Contingencia

4.1 Las Medidas de Asociación 36

4.2 Intepretación de las Razones de Oportunidades y de Riesgo 40

4.3 Prueba de Independencia 42

4.4 Independencia y Análisis de Residuales 51

Literatura Citada 55

Capítulo 1

Suficiencia y Método de Verosimilitud Máxima en el Análisis de Datos Categóricos

Este trabajo trata sobre el análisis de datos categóricos, y el capítulo se dedica a introducir la idea de variable aleatoria categórica, así como a discutir el estadístico suficiente en el análisis de este tipo de datos. Posteriormente, se presenta el método de verosimilitud máxima para obtener estimadores puntuales y probar hipótesis sobre la distribución de una variable categórica.

1. Categorías de datos

Como punto de partida, es conveniente iniciar la presentación con una discusión de la idea de *categoría* de datos, noción que se generó después de analizar la presentación en Agresti (1996, 2004) y Bishop *et. al.*(1975). Considere un experimento aleatorio y sea Ω el conjunto de posibles resultados, también llamado *el espacio muestral*. Por ejemplo, si se observa el tiempo (en segundos) de vida útil T de un objeto, digamos un foco, entonces $\Omega = [0, \infty)$, o si se registra el peso en kilogramos de una persona, entonces $\Omega = [0, 120]$. Como otro ejemplo, suponga que se le pregunta a un estudiante de la Universidad su estado de origen; en este caso, $\Omega = \{\text{Aguascalientes}, \ldots, \text{Zacatecas, Otro}\}$ consiste de los treinta y dos estados de la república, y se ha incluido la posibilidad 'Otro', para cubrir la eventualidad de que algún estudiante encuestado sea extranjero. En ambos casos, las especificaciones de Ω son razonables, en el sentido de que incluyen cada una de las posibilidades que pueden surgir al realizar el experimento. Sin embargo, existe una clara diferencia entre los tres espacios muestrales considerados. En el último caso, el espacio muestral es *finito*, mientras que en los dos primeros ejemplos, el espacio muestral es todo un intervalo y contiene, por lo tanto, un número

infinito de resultados. El análisis de datos categóricos trata, fundamentalmente, con experimentos aleatorios en los que el espacio muestral es finito. A primera vista, este contexto es bastante restrictivo, pues como hemos visto los espacios muestrales con un número infinito de miembros surgen de forma natural en problemas sencillos. Estos casos se incluyen la teoría general por medio de la idea de partición del espacio muestral, la cual se introduce a continuación.

Definición 1.1.1. Considere un espacio muestral Ω arbitrario. Una *partición* $\mathcal{P}$ de Ω es una colección $\mathcal{P} = \{A_1, A_2, \ldots, A_N\}$ de subconjuntos de Ω tales que

$$A_1 \cup A_2 \cup \cdots \cup A_N = \Omega$$

y

$$A_i \cap A_j = \emptyset, \quad i \neq j.$$

Cada uno de los subconjuntos $A_1, A_2, \ldots, A_N$ que pertenecen a $\mathcal{P}$ se llama una *categoría* de datos correspondiente a $\mathcal{P}$.

Note que $\mathcal{P} = \{A_1, A_2, \ldots, A_N\}$ es una partición del espacio muestral cuando y sólo cuando cada elemento de Ω pertenece a exactamente una de las categorías A_i. Como ejemplo, considere de nuevo la medición del tiempo de vida de un foco, experimento para el cual el espacio muestral es $\Omega = [0, \infty)$. En este caso es posible particionar Ω en, digamos, tres categorías,

$$A_1 = [0, 1000), \quad A_2 = [1000, 15000], \quad A_3 = (1500, \infty).$$

Cuando el tiempo de vida T del foco pertenezca a A_1, de manera que $T < 1000$, el foco puede clasificarse como 'defectuoso', si T pertenece a A_2, de modo que $1000 \leq T \leq 1500$, entonces el foco es 'bueno', mientras que el foco es 'excelente' si $T > 1500$. En general, las particiones del espacio muestral surgen cuando los valores de una variable aleatoria se agrupan en categorías, como en este ejemplo. Después de esta discusión, se está en posibilidad de introducir, formalmente, la idea de variable categórica asociada a una partición $\mathcal{P}$.

Definición 1.1.2. Considere una variable aleatoria arbitraria T que toma valores en el espacio muestral Ω, y sea $\mathcal{P} = \{A_1, A_2, \ldots, A_N\}$ un partición de Ω en categorías

$A_1, A_2, \ldots, A_N$. La variable categórica X inducida por T y la partición $\mathcal{P}$ toma valores en el conjunto de catogorías $A_1, A_2, \ldots, A_N$, y se define como

$$X = A_i \iff T \in A_i.$$

De acuerdo a esta especificación, una variable es categórica si sus posibles valores son las categorías de una partición, y por lo tanto, los posibles valores de X forman un conjunto finito. De hecho, es posible definir una variable categórica como una que toma solo un conjunto finito de valores (*i.e.*, una variable discreta), pero ese enfoque no pone de relieve el hecho de que, aún en experimentos con espacio muestral infinito, las variables categóricas surgen a través de la partición del espacio muestral. Después de este preámbulo, a continuación se analiza el problema de estimar la distribución de X.

2. El Estadístico Suficiente

Considere una variable categórica X cuyos valores son las categorías A_1,..., A_k. Las probabilidades de que X asuma cada uno esos posibles valores, las cuales se denotan mediante $\pi_1, \ldots, \pi_k$, son de la mayor importancia:

$$\pi_i = P[X \in A_i], \quad i = 1, , 2, \ldots, k. \tag{2.1}$$

Estas probabilidades contienen toda la información sobre X desde el punto de vista estadístico, y los experimentos que se hagan para observar X tienen como objetivo estimar dichas probabilidades. En esta sección, se considera el siguiente problema: Suponga que se realizan N repeticiones del experimento para obtener, en cada uno de ellos, el valor asumido por la variable X. En este caso, se desea responder a la pregunta ¿qué información debe conservarse para estimar las probabilidades $\pi_j,\ j = 1, 2, \ldots, k$? Para estudiar este cuestionamiento, considere el siguiente contexto. Se tienen

(i) Variables categóricas $X_1, X_2, \ldots, X_N$ cuyos valores pertenecen al conjunto de categorías, $A_1, A_2, \ldots, A_k$;

(ii) $X_1, X_2, \ldots, X_N$ son independientes con distribución común dada por (2.1), esto es,

$$P[X_j = A_i] = \pi_i, \quad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, N.$$

En este caso, sea $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ y $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ un vector cuyas componentes son categorías. Observando que

$$P[X_j = x_j] = \prod_{i=1}^{k} \pi_i^{I[x_j = A_i]},$$

se desprende que la distribución del vector X está dada por

$$\begin{aligned} P[X = \mathbf{x}] &= P[X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N] \\ &= \prod_{i=1}^{k} \pi_i^{I[x_1 = A_i]} \times \prod_{i=1}^{k} \pi_i^{I[x_2 = A_i]} \times \cdots \times \prod_{i=1}^{k} \pi_i^{I[x_N = A_i]} \\ &= \prod_{i=1}^{k} \pi_i^{\sum_{j=1}^{N} I[x_j = A_i]}; \end{aligned} \tag{2.2}$$

denotando mediante $f(\mathbf{x}; \pi)$ a la función de probabilidad de X, esta igualdad establece que

$$f(\mathbf{x}; \pi) = \prod_{i=1}^{k} \pi_i^{\sum_{j=1}^{N} I[x_j = A_i]} \tag{2.3}$$

Esta expresión muestra que $f(\mathbf{x}; \pi)$ depende de $\mathbf{x}$ sólo a través de las cantidades

$$n_i = \sum_{j=1}^{N} I[x_j = A_i], \quad i = 1, 2, \ldots, k;$$

note que n_i es la frecuencia con que la categoría A_i ocurre en el vector $\mathbf{x}$. A partir de esta fórmula, se obtiene el siguiente resultado.

Teorema 1.2.1 Sean $X_1, X_2, \ldots, X_N$ variables categóricas independientes con valores en el conjunto de categorías, $A_1, A_2, \ldots, A_k$ y con distribución común dada por

$$P[X_j = A_i] = \pi_i, \quad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, N,$$

donde el vector de probabilidades $\pi = (\pi_1, \ldots, \pi_k)$ es desconocido. En este caso,

(i) El estadístico

$$S = (N_1, N_2, \ldots, N_k), \quad N_i = \sum_{j=1}^{N} I[X_j = A_i], \quad i = 1, 2, \ldots, k$$

es *suficiente* para π.

(ii) La distribución de S es multinomial con parámetros N y π, es decir, para todos los enteros no negativos $n_1, n_2, \ldots, n_k$ tales que

$$n_1 + n_2 + \cdots + n_k = N,$$

se tiene que

$$P[N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k] = \binom{N}{n_1, n_2, \ldots, n_k} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k}$$

(iii) La media y la varianza de S están dados por

$$E[S] = E \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_k \end{bmatrix} = N \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{bmatrix} = N\pi$$

$$\text{Var}\,(S) = N\left[\text{diag}(\pi) - \pi\pi'\right],$$

donde

$$\text{diag}(\pi) = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_k \end{bmatrix}$$

es la matriz diagonal con las componentes de π a lo largo de la diagonal principal.

Demostración. La parte (i) se desprende, via (2.3), del criterio de factorización de Fisher; vea, por ejemplo, Ferguson (1967), o Kendall y Stuart (1979). A partir de este punto, las partes (ii) y (iii) son consecuencia de la construcción básica de la distribución multinomial y de sus propiedades (Rao, 1973). □

El significado práctico del Teorema 1.2.1(i) es que, al investigar la distribución de una variable categórica y realizar repeticiones independientes del experimento que la genera, los datos que deben conservarse son las frecuencias N_i con que se observa que el resultado pertenece a la catagoría A_i. Por esta razón, el material fundamental de análisis se presenta en la forma de tablas, como la siguiente:

A_1	A_2	$\cdots$	A_k	
N_1	N_2	$\cdots$	N_k	N

Tabla Básica de Frecuencias Para Datos Categóricos

Aquí, N_i es la frecuencia observada de la categoría i, mientras que N es el número total de observaciones (repeticiones del experimento); note que N_i es el valor de una variable aleatoria,, y que N es el número total de datos, usualmente fijo. Por supuesto,

$$N = N_1 + N_2 + \cdots + N_k.$$

A continuación se aborda el problema de estimar la distribución de una variable categórica.

3. Método de Verosimilitud Máxima

Considere el estadístico suficiente $S = (N_1, N_2, \ldots, N_k)'$ en el Teorema 1.2.1. Para cada vector $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ de enteros no negativos que sumen N, se tiene que

$$f(\mathbf{n}, \pi) = P[N_1 = n_1, \ldots, N_k = n_k] = \binom{N}{n_1, n_2, \ldots, n_k} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k},$$

de manera que la verosimilitud asociada al evento $[N_1 = n_1, \ldots, N_k = n_k]$ es

$$L(\pi; \mathbf{n}) = \binom{N}{n_1, n_2, \ldots, n_k} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k}, \tag{3.1}$$

y el método de verosimiltud máxima prescribe estimar π por el vector que maximiza esta función. Para encontrar el maximizador, es conveniente tomar el logaritmo natural para obtener

$$\mathcal{L}(\pi, \mathbf{n}) = \log(L(\pi; \mathbf{n})) = \log \binom{N}{n_1, n_2, \ldots, n_k} + \sum_{i=1}^{k} n_i \log(\pi_i)$$

de donde se desprende que

$$\partial_{\pi_i} \mathcal{L}(\pi, \mathbf{n}) = \frac{n_i}{\pi_i}, \quad i = 1, 2, \ldots, k.$$

Recordando que $\pi_1 + \pi_2 + \cdots + \pi_k = 1$, después de observar que

$$\partial_{\pi_i} [\pi_1 + \pi_2 + \cdots + \pi_k] = 1,$$

el método de multiplicadores de Lagrange (Khuri, 2002), estipula que el maximizador debe satisfacer $\partial_{\pi_i}\mathcal{L}(\pi, \mathbf{n}) = c \times 1 = c, \quad i = 1, 2, \ldots, k$ para alguna constante c, esto es,

$$\frac{n_i}{\pi_i} = c, \quad i = 1, 2, \ldots, k.$$

De aquí se desprende que

$$\pi_i = \frac{n_i}{c} \tag{3.2}$$

y entonces

$$1 = \sum_{i=1}^{k} \pi_i = \sum_{i=1}^{k} \frac{n_i}{c} = \frac{\sum_{i=1}^{k} n_i}{c} = \frac{N}{c}$$

esto es, $c = N$, igualdad que combinada con (3.2) conduce a $\pi = n_i/N$ para todo i. Esto demuestra la primera parte del siguiente resultado.

Teorema 1.3.1 (i) El estimador de verosimilitud máxima de π, denotado por

$$\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k),$$

está dado por

$$\hat{\pi}_i = \frac{N_i}{N}, \quad i = 1, 2, \ldots, k.$$

(ii) El estimador $\hat{\pi}$ es insesgado, es decir,

$$E[\hat{\pi}] = \pi,$$

mientras que

$$\text{Var}\,(\hat{\pi}) = \frac{1}{N}[\text{diag}(\pi) - \pi\pi'].$$

Demostración. Como se mencionó anteriormente, la parte (i) se desprende del argumento previo al enunciado del teorema, mientras que la parte (ii) se sigue observando que $\hat{\pi} = (N_1, N_2, \ldots, N_k)'/N$, y combinando el Teorema 1.2.1 con las propiedades estándar del valor esperado y la varianza establecidas, por ejemplo, en Rao (1973) o Ferguson (1967). □

A continuación se aborda el problema de prueba de una hipótesis simple en el análisis de datos categóricos.

4. Prueba de Razón de Verosimilitud

En esta sección se considera el problema de probar la hipótesis

$$\mathcal{H}_0: \pi = \pi_0 \tag{4.1}$$

donde

$$\pi_0 = (\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{k,0})'$$

es un vector de probabilidad fijo. Note que la verosimilitud el vector π_0 es

$$L(\pi_0; \mathbf{n}) = \binom{N}{n_1, n_2, \ldots, n_k} \pi_{0,1}^{n_1} \pi_{0,2}^{n_2} \ldots \pi_{0,k}^{n_k}$$

Por otro lado, la mayor verosimilitud posible es

$$L(\hat{\pi}; \mathbf{n}) = \binom{N}{n_1, n_2, \ldots, n_k} \hat{\pi}_1^{n_1} \hat{\pi}_2^{n_2} \ldots \hat{\pi}_k^{n_k}$$

por lo que el cociente de ambas verosimiltudes es

$$\begin{aligned} \Lambda(\mathbf{n}) &= \frac{L(\hat{\pi}; \mathbf{n})}{L(\pi_0; \mathbf{n})} \\ &= \frac{\hat{\pi}_1^{n_1} \hat{\pi}_2^{n_2} \ldots \hat{\pi}_k^{n_k}}{\pi_{0,1}^{n_1} \pi_{0,2}^{n_2} \ldots \pi_{0,k}^{n_k}} \\ &= \left(\frac{\hat{\pi}_1}{\pi_{0,1}}\right)^{n_1} \left(\frac{\hat{\pi}_2}{\pi_{0,2}}\right)^{n_2} \cdots \left(\frac{\hat{\pi}_k}{\pi_{0,k}}\right)^{n_k}; \end{aligned} \tag{4.2}$$

observe que este cociente es mayor que 1. Defina ahora el estadístico G^2 mediante

$$G^2 = 2\log(\Lambda(\mathbf{n})) = 2\sum_{i=1}^{n} n_i \log\left(\frac{\hat{\pi}_i}{\pi_{0,i}}\right)$$

el cual es no negativo. La teoría general del método de verosimiltud máxima presentada en Ferguson (1967), establece la siguiente conclusion:

Teorema 1.4.1 Conforme N crece, la distribución de G^2 converge a una distribución ji-cuadrada con $k-1$ grados de libertad, es decir,

$$G^2 \xrightarrow{\text{d}} \chi^2_{k-1}.$$

Note que, bajo el supuesto de que $\mathcal{H}_0$ es cierta, el número esperado de observaciones en la categoría i es

$$\mu_i = NP[X = A_i] = N\pi_{0,i}. \tag{4.3}$$

Por otro lado, la frecuencia real observada de la categoría i es $n_i = N\hat{\pi}_i$, por lo que se tiene

$$\frac{\hat{\pi}_i}{\pi_{0,i}} = \frac{N\hat{\pi}_i}{N\pi_{0,i}} = \frac{n_i}{\mu_i}$$

y por lo tanto, el estadístico G^2 puede escribirse como

$$G^2 = 2\log(\Lambda(\mathbf{n})) = 2\sum_{i=1}^{n} n_i \log\left(\frac{n_i}{\mu_i}\right) \tag{4.4}$$

Cuando cada frecuencia observada n_i es similar a la correspondiente frecuencia esperada μ_i bajo la hipótesis $\mathcal{H}_0$—de manera que los datos son consistentes con el supuesto de que $\mathcal{H}_0$ es cierta—se tiene que n_i/μ_i es cercano a 1, y entonces $\log(n_i/\mu_i)$ es cercano a cero, dando como resultado que G^2 tome valores moderados. De esta manera, valores grandes de G^2 constituyen evidencia de que $\mathcal{H}_0$ no es cierta. Esto sugiere el procedimiento de prueba de la hipótesis $\mathcal{H}_0$ delineado en el siguiente lema (Goodman y Kruskal, 1979, Greenwood y Nikulin, 1996, Read y Cressie, 1988).

Lema 1.4.1 Dado un nivel de significancia $\alpha < 1$, considere la hipótesis $\mathcal{H}_0$ en (4.1), y sea G^2 el estadístico en (4.4). En este caso, el procedimiento de prueba que rechaza $\mathcal{H}_0$ cuando

$$G^2 > \chi^2_{k-1,\alpha}$$

tiene un nivel aproximado de significancia α, y la aproximación es mejor a medida que el número N de datos crece.

La prueba delineada en este lema es la 'prueba de razón de verosimilitud'y fue propuesta originalmente en (Wald, 1943).

5. Prueba Ji-Cuadrada

Otro procedimento para probar $\mathcal{H}_0$ es la prueba ji-cuadrada, la cual utiliza el estadístico

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - \mu_i)^2}{\mu_i} \tag{5.1}$$

Note que si la frecuencia observada n_i es similar a la correspondiente frecuencia esperada μ_i, entonces $(n_i - \mu_i)^2$ no será grande, y por lo tanto X^2 tomará valores moderados. De esta manera, valores 'grandes' de X^2 son evidencia contra $\mathcal{H}_0$. Por otro lado, no es difícil demostrar que, bajo $\mathcal{H}_0$, $G^2 - X^2$ converge a cero conforme el número de datos aumenta, esto es,

$$G^2 - X^2 \xrightarrow{\text{P}} 0,$$

y por lo tanto, la distribución de X^2 se aproxima a χ^2_{k-1} al aumentar N (vea el Teorema 1.4.1):

$$X^2 \xrightarrow{\text{d}} \chi^2_{k-1} \quad \text{conforme } N \to \infty.$$

Este argumento sugiere el procedimiento de prueba en el siguiente lema (Agresti, 1996, 2004, Goodman y Kruskal, 1979).

Lema 1.5.1 Dado un nivel de significancia $\alpha < 1$, considere la hipótesis $\mathcal{H}_0$ en (4.1), y sea X^2 el estadístico en (5.1). En este caso, el procedimiento de prueba que rechaza $\mathcal{H}_0$ cuando

$$X^2 > \chi^2_{k-1,\alpha}$$

tiene un nivel aproximado de significancia α, y la aproximación es mejor a medida que el número N de datos crece.

Ambas pruebas, la ji-cuadrada en este lema, y la G^2 en el Lema 1.4.1, conducen a conclusiones similares siempre que el tamaño de la muestra N sea suficientemente grande.

Después de introducir las ideas básicas anteriores, éstas se ilustrarán con una serie de ejemplos en el siguiente capítulo.

Capítulo 2

Ejemplos Sobre las Ideas Básicas

En este capítulo se presentan ejemplos detallados sobre las ideas presentadas hasta ahora y sobre otras nociones básicas, incluyendo las distribuciones Binomial y de Poisson. La presentación incluye la distinción entre variables de respuesta y variables explicatorias, y es conveniente tener una idea clara sobre la distinción entre ellas. En general, un experimento estadístico se realiza para obtener el valor de una o más variables de respuesta y averiguar la manera que otras variables, llamadas explicatorias, influyen o se asocian con las primeras. De manera intuitiva, una variable involucrada en un experimento estadístico es de respuesta, si la obtención de su valor es el propósito básico de la realización del experimento, mientras que las restantes variables son explicatorias; comúnmente, las variables de respuesta se observan al final del experimento, mientras que las variables explicatorias son observadas o fijadas por el analista antes de concluir el experimento.

1. Variables de Respuesta y Explicatorias

Las ideas de variable de respuesta y variable explicatoria, así como de variable ordinal y nominal son discutidas a continuación.

Ejemplo 1.1. En los siguientes casos, identifique las variable de respuesta y las variables explicatorias.

(a) Actitud hacia el control de armas (a favor, opuesto), Género (femenino, masculino), nivel de educacion de la madre (preparatoria, universidad).

(b) Enfermedad cardiaca (si, no), Presión arterial, Nivel de Colesterol.

(c) Raza (Blanca, no blanca), Religión (católica, judía, protestante), Voto presidencial (demócrata, republicano, otro), Ingreso anual.

Solución. (a) El propósito del experimento es averiguar la actitud hacia el control de armas de las personas entrevistadas; la variable de respuesta es entonces 'Actitud hacia el control de armas'. Las otras variables, Género y Nivel de educación de la madre, serán observadas o fijadas por el entrevistador antes de obtener la variable de respuesta, y son, por lo tanto, variables explicatorias.

(b) En este contexto, el experimento subyacente consiste en seleccionar varias personas y averiguar si tienen o no enfermedad cardiaca, por lo que la variable de respuesta es 'Presencia de enfermedad cardiaca'; las otras variables, Presión arterial y Nivel de colesterol, son explicatorias y son registradas por el analista para investigar la asociación que tienen con la variable de respuesta.

(c) En este caso, el experimento que se sugiere consiste en seleccionar varias personas y preguntarles cual será su decisión en la próxima elección presidencial. La variable de respuesta es entonces 'Voto presidencial', mientras que las variables explicatorias son Religión, Raza, e Ingreso anual. El analista tratará, después de la realización de las entrevistas y la obtención de datos, de averiguar la relación entre las variables explicatorias y la variable de respuesta. □

Ejemplo 1.2. ¿ Cuál escala de medición—nominal u ordinal—es más apropiada para las siguientes variables?

(a) Afiliación política partidaria (demócrata, republicano, no afiliado).

(b) Máximo grado académico obtenido (doctorado, maestría, licenciatura, preparatoria, otro, ninguno)

(c) Estado clínico de un paciente (bueno, regular, serio, crítico)

(d) Ubicación de un centro de salud.

(e) Bebida favorita(cerveza, jugo, leche, vino, otra)

Solución. (a) Los posibles valores de la variable categórica 'Afiliación política'—demócrata, republicano, y no afiliado–no tienen un orden natural entre ellos. Por la tanto, la variable categórica es nominal.

(b) Los diversos valores que la variable categórica 'Máximo grado académico obtenido' puede tomar están naturalmente ordenados, por lo que la variable es ordinal.

(c) Los posibles valores de la variable categórica 'Estado Clínico' están naturalmente ordenados, por lo que la variable es ordinal.

(d) La ubicación de un hospital, cuyos valores son localidades—como Arteaga, Zitácuaro, Saltillo norte, Monterrey—no están ordenados, de manera que la variable 'Ubicación de la Clínica' es nominal. □

2. Ejemplos sobre las Distribuciones Binomial y de Poisson

Los siguientes ejemplos se refieren a las distribuciones binomial, geométrica y de Poisson. Los detalles de la teoría correspondiente se pueden encontrar en Brown *et. al.* (2001), o en Fleiss (1981).

Ejemplo 2.1. Las obleas de silicón fabricadas por una compañía tienen un promedio de defectos de 1.0 por oblea. Si el número de defectos tiene la distribución de Poisson, encuentre la probabilidad de que una oblea tenga (a) 0 defectos, (b) un defecto, y (c) al menos dos defectos.

Solución. Denotando por N el número aleatorio de defectos que una oblea presenta, se tiene que

$$P[N = n] = e^{-\mu}\frac{\mu^n}{n!}, n = 0, 1, 2, 3, \ldots$$

donde μ es el número esperado de defectos en una oblea; en este caso, $\mu = 1.0$

(a) $P[N = 0]e^{-\mu}\mu^0/0! = e^{-\mu} = e^{-1}$:

(b) $P[N = 1]e^{-\mu}\mu/1! = e^{-1}1/1! = e^{-1}$:

(c) $P[N \geq 2] = 1 - P[N = 0] - P[N = 1] = 1 - e^{-1} - e^{-1} = 1 - 2e^{-1}$. □

Ejemplo 2.2. Cada una de las 100 preguntas de respuesta múltiple de un examen tiene cuatro respuestas de las cuales sólo una es correcta. Para cada pregunta, un estudiante selecciona aleatoriamente una de las cuatro respuestas y la marca como correcta. Especifique la distribución del número de respuestas correctas que el estudiante tiene en el examen, y con base en el el valor esperado y la varianza de la distribución ¿sería sorprendente si el estudiante tiene al menos 50 respuestas correctas?

Solución. Cada vez que el estudiante observa una pregunta y selecciona una de las cuatro respuestas posibles como correcta, esta realizando un ensayo de Bernoulli. Ocurre un ''exito' si la respuesta marcada es efectivamente correcta, y por lo tanto la probabilidad de exito es $\pi = 1/4$, pues la respuesta marcada se elige al azar entre las cuatro disponibles y sólo una es correcta. Como este proceso se repite para cada una de las $n = 100$ preguntas, suponiendo la independencia entre las selecciones hechas en cada caso, se tiene que el número total de éxitos N es una variable aleatoria con distribución binomial de parámetros $n = 100$ y $\pi = 1/4$, esto es,

$$\begin{aligned} P[N = n] &= \binom{100}{n} \pi^n (1-\pi)^{100-n} \\ &= \binom{100}{n} \left(\frac{1}{4}\right)^n \left(\frac{3}{4}\right)^{100-n}, \quad n = 0, 1, 2, \ldots, 100. \end{aligned}$$

El valor esperado de N y la desviación estándar de N son, $\mu = n\pi = 25$ y $\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{75} \approx 8.6$. Por lo tanto, una observación de por lo menos 50 respuestas correctas ('éxitos'), esto es $N \geq 50$, corresponde a un valor estandarizado

$$\frac{N - 25}{\sqrt{75}} \geq \frac{50 - 25}{\sqrt{75}} \approx 3$$

Por el Teorema central de límite, $\dfrac{N - 25}{\sqrt{75}}$ tiene, aproximadamente, distribución normal estándar, y por lo tanto

$$P[N \geq 50] = P\left[\frac{N - 25}{\sqrt{75}} \geq \frac{50 - 25}{\sqrt{75}}\right] \approx P\left[\frac{N - 25}{\sqrt{75}} \geq 3\right] \approx .001.$$

En consecuencia, observar $N \geq 50$ cuando las respuestas a las preguntas se eligen al azar, es un evento sumamanete extraño, y al tener lugar, debe conducir a la conclusión de que las respuestas no fueron elegidas aleatoriamente. □

Ejemplo 2.3. Una moneda balanceada se lanza dos veces; sea Y el número de 'águilas' obtenidas.

(a) Especifique las probabilidades para los posibles valores de Y, e indique su distribución.

(b) Calcule las probabilidades para la distribución de Poisson que tiene la misma media que la distribución en el inciso (a). ¿Cómo se comparan su varianza con aquélla del inciso (a)?

(c) Para cada lanzamiento de una moneda posiblemente no balanceada, denote mediante π a la probabilidad de obtener un 'águila'. Suponga que hay cero águilas en dos lanzamientos. Encuentre el estimador de verosimilitud máxima de π. ¿Es razonable este valor? (El método de Bayes, una alternativa que combina las creencias previas acerca del parámetro con los datos muestrales, proporciona la estimación no nula $\hat{\pi} = (Y+1)/(N+2) = (0+1)/(2+2) = .25$ cuando, previamente al experimento, se cree que π tiene distribución uniforme en $(0,1)$.)

Solución. (a) Cada lanzamiento de la moneda representa un ensayo de Bernoulli, en el que la probabilidad de 'éxito'—obtener un águila—es $p = 1/2$, pues la moneda es balanceada. Como se realizan dos lanzamientos, el total de éxitos Y tiene distribución binomial con parámetros $n = 2$ y $\pi = 1/2$,, esto es,

$$P[Y = y] = \binom{2}{y}\left(\frac{1}{2}\right)^{y}\left(1-\frac{1}{2}\right)^{2-y} = \frac{1}{4}\binom{2}{y}, \quad y = 0, 1, 2.$$

La esperanza y varianza de Y son $E[Y] = 2\pi = 1$ y $\text{Var}\,(Y) = n\pi(1-\pi) = 1/2$, respectivamente.

(b) Para una variable Y_P con distribución de Poison con media 1, se tiene que $P[Y_P = y] = e^{-1}(1)^y/y! = e^{-1}/y!$ para $y = 0, 1, 2, \ldots$. La siguiente tabla muestra explícitamente los valores de las probabilidades:

	$Y \sim B(2, 1/2)$	$Y_P \sim \mathcal{P}(1)$
y	$P[Y = y] = \frac{1}{4}\binom{2}{y}$	$P[Y_P = y] = e^{-1}/y!$
0	0.25	0.3679
1	0.50	0.3679
2	0.25	0.1840
3	0.00	0.0613
4	0.00	0.0153
5	0.00	0.0030
6	0.00	0.0005

(c) La función de verosimilitud asociada a la observación $Y = 0$, es

$$L(\pi; 0) = \binom{2}{0}\pi^0(1-\pi)^{2-0} = (1-\pi)^2, \quad \pi \in [0, 1]$$

Esta función es decreciente, y por lo tanto alcanza su valor máximo cuando π asume el valor 0. Por lo tanto, el estimador de verosimilitud máxima es $\hat{\pi} = 0$. Este valor del estimador es con frecuencia cuestionado, pues con solo observar dos lanzamientos de la moneda, al no obtener águilas estipula que la probabilidad π de un águila en un lanzamiento arbitrario es cero. Sin embargo, debe recordarse que un estimador es sólo una aproximación al verdadero valor del parámetro desconocido—en este caso π—y que los valores de un estimador comúnmente no coinciden con el valor correcto del parámetro; en general, π y cualquier estimador $\hat{\pi}$ difieren, debido a las fluctuaciones aleatorias. El caso que nos ocupa no es más que un recordatorio de este hecho.

Ejemplo 2.4. Con relación al ejemplo previo,

(a) Calcule las probabilidades binomiales cuando el número de ensayos es $N = 2$ y la probabilidad de obtener un águila es (i) $\pi = 0.6$ y (ii) $\pi = 0.4$

(b) Suponga que se observa $Y = 1$. Determine la función de verosimilitud,

(c) Muestre que la estimación de verosimilitud máxima de π es $\hat{\pi} = 0.5$.

Solución. (a) Las probabilidades requeridas se muestran en la siguiente tabla.

	$Y \sim B(2, 0.6)$	$Y \sim B)2, 0.4)$
y	$P[Y = y] = \binom{2}{y}(0.6)^y(0.4)^{2-y}$	$P[Y = y] = \binom{2}{y}(0.4)^y(0.6)^{2-y}$
0	0.16	0.36
1	0.48	0.48
2	0.36	0.16

Esta tabla es una muestra del siguiente hecho general: Si $Y \sim B(n, \pi)$, entonces $n - Y \sim B(n, 1 - \pi)$.

(b) Cuando $Y \sim B(2, \pi)$, la función de verosimilitud asociada al dato $Y = 1$ es

$$L(\pi; 1) = P_\pi[Y = 1] = \binom{2}{1}\pi^1(1 - \pi)^1 = \pi(1 - \pi), \quad \pi \in [0, 1].$$

(c) Note que $L(\pi; 1) = \pi(1-\pi)$ es una función cuadrática para la cual el coeficiente de π^2 es -1, y por lo tanto alcanza su valor máximo en el punto en el que la

derivada se anula. Como $D_\pi L(\pi; 1) = 1 - 2\pi$, la derivada se anula en el punto $\pi = 0.5$, y por lo tanto $\hat{\pi} = 0.5$. □

Ejemplo 2.5. En su autobiografía—*A Sort of Life*—el autor británico Graham Greene describe un período de depresión severa que sufrió durante su vida, en el cual jugó a la ruleta rusa, 'juego' que consiste en colocar una bala en una de las seis cámaras de un revólver, girar la recámara del arma, y finalmente dispararse a la cabeza una vez.

(a) Greene jugó a la ruleta rusa seis veces, y tuvo la suerte de que en ninguna ocasión el arma disparó la bala. Encuentre la probabilidad de ese resultado.

(b) Suponga que una persona juega a la ruleta rusa una y otra vez hasta que la bala se dispara en el juego Y. Muestre que la probabilidad de que $Y = y$ es $(5/6)^{y-1}(1/6)$ para $y = 1, 2, 3, \ldots$, la cual se conoce como distribución geométrica.

Solución. (a) Cada vez que se juega a la ruleta rusa, se realiza un experimento de Bernoulli, en el que la probabilidad de 'éxito'—identificado con el evento de que la bala se dispara—es $1/6$, pues un revólver tiene seis cámaras, y la bala se coloca al azar en alguna de ellas. Al observar seis repeticiones independientes del juego, el número total de éxitos es una variable aletoria N con distribución $N \sim B(6, 1/6)$, esto es,

$$P[N = n] = \binom{6}{n} \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{6-n}, \quad n = 0, 1, 2, \ldots, 6.$$

Greene tuvo la suerte de observar $N = 0$, evento que tiene probabilidad $(5/6)^6 = 0.1389$.

(b) Denote mediante F_i el evento 'falla en el i-ésimo juego', esto es, la bala no se dispara en la i-ésima repetición del juego de la ruleta rusa. Con esta notación, los eventos F_i son independientes y $P[F_i] = 5/6$ para cada i. Por otro lado, el evento $Y = y$ se expresa en términos de los eventos F_i como

$$[Y = y] = F_1 \cap F_2 \cap \cdots \cap F_{y-1} \cap F_y^c;$$

note que el lado derecho ocurre cuando los juegos $1, 2, \ldots, y-1$ resultan en que la bala no se dispara, mientras que en el ensayo y la bala si se dispara, precisamente

lo que significa $Y = y$. Por la independencia de los eventos F_i, se tiene que

$$\begin{aligned} P[Y = y] &= P[F_1] \times P[F_2] \times \cdots \times P[F_{y-1}] \times P[F_y^c] \\ &= (5/6) \times (5/6) \times \cdots \times (5/6) \times (1/6) = (5/6)^{y-1}(1/6), \end{aligned}$$

concluyendo el argumento. □

3. Prueba de Hipótesis para un Parámetro Binomial

En esta sección se abordan problemas relacionados con la estimación de proporciones, particularmente la construcción de intervalos de confianza por medio del teorema central de límite (Brown *et. al.*, 2001).

Ejemplo 3.1. Una muestra de pacientes con dolor muscular ha estado tomando un analgésico para disminuir el dolor. Se afirma que un nuevo analgésico produce mayor alivio que el estándar actualmente utilizado. Después de probar el nuevo producto, 40 pacientes reportaron mayor alivio con el analgésico estándard, mientras que 60 reportaron mayor alivio con el nuevo medicamento.

(a) Pruebe, al nivel de significancia de 0.05%, la hipótesis de que la probablidad de mayor alivio es la misma para ambos analgésicos, el nuevo y el estándar.

(b) Construya un intervalo de 95% de confianza para la probabilidad de mayor alivio con el nuevo analgésico.

Solución. En este ejemplo, a 100 pacientes se les administraron ambos analgésicos durante algún período, al final del cual se le preguntó a cada uno cual de los dos analgésicos le produjo mayor alivio. Defina la variable aleatoria Y_i como 1, si el nuevo analgésico es considerado mejor por el i-ésimo paciente, mientras que $Y_i = 0$ en caso contrario. En este caso, $Y_1, \ldots, Y_{100}$ son variables de Bernoulli y la probabilidad de que tomen el valor 1 es π, la probabilidad de que el nuevo analgésico produzca mayor alivio. Suponiendo que las respuestas sean independientes, se tiene que $Y = \sum_{i=1}^{100} Y_i$—el número de pacientes que reportan al nuevo analgésico como el mejor—tiene distribución binomial con parámetros $n = 100$ y probabilidad de éxito π. De acuerdo a los datos, se observó $Y = 60$, de manera que el esimador de verosimilitud máxima es

$$\hat{\pi} = \frac{Y}{100} = \frac{60}{100} = 0.60$$

(a) La hipótesis en cuestión es $\mathcal{H}_0: \pi = \pi_0 = 0.5$, y el estadístico de prueba es

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/100}} = \frac{0.6 - 0.5}{\sqrt{0.5(1-0.5)/100}} = 2;$$

el cual tiene, aproximadamente, distribución normal estándar. Como el percentil (bilateral) de orden 0.05% de esta distribución límite es

$$z_{0.025} = 1.96,$$

observando que $2 = |z| > z_{0.025} = 1.96$, se rechaza la hipótesis $\mathcal{H}_0$ al nivel de significancia de 0.05%.

(b) El intervalo de 0.95% de confianza para π es

$$\hat{\pi} - z_{0.025}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{100}} \leq \pi \leq \hat{\pi} + z_{0.025}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{100}}$$

El significado de este enunciado es que en el 95% de los experimentos que se realicen las anteriores desigualdades son válidas. Por lo tanto, en el experimento actual que se ha realizado, el analista confia en que las desigualdes que resulten son correctas; al usar este procedimiento un gran numero de veces, en el 95% de los casos se obtendran conclusiones ciertas, y en el 5% de las ocasiones las conclusiones serán equivocadas. Al sustituir los datos observados, se obtiene

$$0.60 - (1.96)\sqrt{\frac{0.60(1-0.60)}{100}} \leq \pi \leq 0.60 + (1.96)\sqrt{\frac{0.60(1-0.60)}{100}},$$

lo cual equivale a

$$0.60 - (1.96)\frac{\sqrt{0.24}}{10} \leq \pi \leq 0.60 + (1.96)\frac{\sqrt{0.24}}{10},$$

de donde se obtiene

$$0.60 - 0.0960 \leq \pi \leq 0.60 + 0.0960,$$

esto es,

$$0.5040 \leq \pi \leq 0.6960.$$

Note que este intervalo de 95% de confianza para π no contiene al valor $\pi_0 = 0.50$, lo cual concuerda con el rechazo de la hipótesis $\mathcal{H}_0: \pi = \pi_0$ al nivel de significancia de 5%. □

Ejemplo 3.2. Con relación al ejemplo precedente, los investigadores desean determinar un tamaño de muestra lo suficientemente grande para lograr estimar la probabilidad π de preferir el nuevo analgésico con un error máximo de 0.08 con confianza de 95%. ¿Qué tan grande debe seleccionarse la muestra para lograr ese objetivo si la verdadera probabilidad π es 0.75?

Solución. Recuerde que la fórmula para un intervalo de $(1-\alpha)\times 100\%$ de confianza para π es

$$\hat{\pi} - z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{3.1}$$

A partir de este intervalo, se desprende que el analista tiene una confianza de $(1-\alpha)\times 100\%$ en que

$$|\hat{\pi} - \pi| \leq z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},$$

esto es, en que la la máxima discrepancia absoluta entre $\hat{\pi}$ y el verdadero valor del parámetro π es $z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$. Como el analista desea un máximo error de estimación de 0.08, entonces debe tenerse que

$$z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.08, \tag{3.2}$$

y a partir de esta ecuación debe determinarse n, el tamaño de la muestra. Con este propósito, recuerde que el nivel de confianza deseado es de $95\% = (1-\alpha)\times 100\%$, de manera que $\alpha = 0.5$ y por lo tanto, $z_{\alpha/2} = z_{0.025} = 1.96$, por lo que la ecuación (3.2) se reduce a $1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.08$, esto es,

$$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \frac{0.08}{1.96}.$$

Al intentar resolver esta ecuación para n aparece una dificultad, a saber, $\hat{\pi}$ es una variable aleatoria. Más aún, no es posible usar un valor observado para $\hat{\pi}$, ya que estos cálculos se hacen en la fase planeación del experimento, antes de realizarlo, y por lo tanto no hay valores observados disponibles. Para solventar esta dificultad, recuerde que, por la ley de los grandes números, $\hat{\pi}$ converge a π_0, el verdadero valor del parámetro conforme el tamaño de la muestra crece, de tal suerte que $\hat{\pi}(1-\hat{\pi}) \approx \pi_0(1-\pi_0)$, de manera que la ecuación anterior estipula que

$$\sqrt{\frac{\pi_0(1-\pi_0)}{n}} \approx \frac{0.08}{1.96}.$$

Recordando que, de acuerdo a los datos del problema, $\pi_0 = 0.75$, se obtiene que

$$n \approx \frac{\pi_0(1-\pi_0)}{(0.08/1.96)^2} = \frac{0.75(0.25)}{(0.08/1.96)^2} = (0.1875)(24.5)^2 = 449.93$$

y entonces, seleccionando el tamaño de la muestra $n \geq 450$. se alcanza el objetivo deseado. □

Ejemplo 3.3. La revista *Newsweek* reportó, en su edición del 27 de marzo de 1989, los resultados de una encuesta realizada por la organización *Gallup* acerca de creencias religiosas. De 750 adultos entrevistados, 24% se declaró creyente en la reencarnación. Tratando a los entrevistados como una muestra aleatoria de la población de adultos, construya un intervalo de 95% de confianza para la proporción de adultos que creen en la reencarnación.

Solución. Denotando por π a la proporción de adultos que creen en la reencarnación, se tiene que el intervalo de 95% de confianza para π es

$$\hat{\pi} - 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}};$$

vea (3.1) teniendo en mente que $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$. En este problema el tamaño de la muestra es $n = 750$ y $\hat{\pi} = 0.24$, de manera que $1.96\sqrt{\hat{\pi}(1-\hat{\pi})/750} = 1.96\sqrt{0.0002432} = 1.96(0.0156) = 0.0306$, y entonces el intervalo de confianza para π es

$$0.24 - 0306 \leq \pi \leq 0.24 + 0306,$$

esto es, $0.20094 \leq \pi \leq 0.2706$. □

Ejemplo 3.4. Un criminalista desea estimar la proporción de ciudadanos que tiene armas de fuego en su casa. En la encuesta nacional de 1991, se preguntó a los encuestados, "¿Tiene usted algún arma de fuego en su casa?"De los encuestados que respondieron a la pregunta, 393 contestaron "si"y 583 dijeron "no". Construya un intervalo de 95% de confianza para la verdadera proporción de personas que tienen armas en su casa.

Solución. Como en el ejemplo anterior, denotando mediante π a la proporción de personas que tienen armas en su casa, se tiene que el intervalo de 95% de confianza para π es

$$\hat{\pi} - 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},$$

donde $\hat{\pi}$ es la proporción observada de respuestas "si" entre las $n = 393+583 = 976$ respuestas registradas, de manera que $\hat{\pi} = 393/976 = 0.4027$. Después de hacer las sustituciones correspondientes en la fórmula anterior, se obtiene

$$0.3719 \leq \pi \leq 0.4334$$

el cual es el intervalo de 95% de confianza para π, de acuerdo al cual la proporción de personas que tienen armas de fuego en su casa se ubica entre 37.19% y 43.34%. Antes de concluir, es conveniente notar un aspecto de la solución. Primeramente, en una encuesta es posible que al ser interrogado sobre una pregunta comprometedora, como es el caso en este problema, algunas personas no contesten, y por lo tanto, no pueda saberse si tienen armas en su casa o no. Por lo tanto, la proporción de personas que respondieron "si"calculada anteriormente , esto es $\hat{\pi} = 0.4027$, es realmente *una estimación de la proporción de personas que tienen armas en su casa dentro de la población de adultos que están dispuestos a responder a la pregunta*, y no una estimación de la proporción de adultos con armas en su casa dentro de toda la población de interés. Este aspecto ilustra una de las inevitables dificultades en el análisis de una encuesta: Estrictamente, las inferencias que se realicen se refieren a la subpoblación de personas que están dispuestas a contestar, y no a toda la población de interés. □

Ejemplo 3.5. Denote mediante $\sigma(Z)$ a la desviación estándar de una variable aletoria Z, esto es, $\sigma(Z) = \sqrt{\text{Var}\,(Z)}$. Si Y es una variable aleatoria y c es una constante positiva, entonces la desviación estándar de cY es $\sigma(cY) = c\sigma(Y)$. Si Y es una variable aleatoria con distribución binomial con parámetros N y π, muestre que $\hat{\pi} = Y/N$ tiene desviación estándar dada por $\sigma(\hat{\pi}) = \sqrt{\pi(1-\pi)/N}$, y utilice esta fórmula para explicar por qué es más fácil estimar π cuando este número está cerca de 0 o 1 que cuando está cercano a 1/2.

Solución. Recuerde que si Y tiene distribución binomial con parámetros N y π, entonces se tiene que $\text{Var}\,(Y) = N\pi(1-\pi)$, de manera que $\sigma(Y) = \sqrt{\text{Var}\,(Y)} =$

$\sqrt{N\pi(1-\pi)}$, Puesto que $\hat{\pi} = Y/N$, usando la relación $\sigma(cY) = c\sigma(Y)$ con $c = 1/N$ se desprende que $\sigma(\hat{\pi}) = \sigma(Y/N) = \sigma(Y)/N = \sqrt{N\pi(1-\pi)}/N$, de donde se desprende que

$$\sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{N}}.$$

A partir de esta fórmula, note que, manteniendo fijo el tamaño de la muestra N, el valor de $\sigma(\hat{\pi})$ se aproxima a cero conforme el parámetro π se acerca a 0 o a 1, mientras que si π se ubica alrededor de $1/2$, entonces $\sigma(\hat{\pi})$ asume un valor cercano a $\sqrt{0.5(1-0.5)/N} = 1/[2\sqrt{N}]$, el cual es positivo. Esto es, el error estándar de $\hat{\pi}$ es mayor cuando π se ubica alrededor de $1/2$ que cuando π se encuentra cercano a cero o a uno; como $\sigma(\hat{\pi})$ es una medida de la precisión del estimador $\hat{\pi}$, se concluye que, para un mismo tamaño de muestra, $\hat{\pi}$ es más preciso—esto es, tiende a estar más cerca del parámetro π—cuando π es cercano a cero o a uno, que cuando π se encuentra alrededor de $1/2$. □

4. Ejemplos sobre la Función de Verosimilitud

A continuación se estudian, de forma detallada, ejemplos referentes a la función de verosimilitud, especialmente el caso de la distribución de Poisson y de la distribución binomial.

Ejemplo 4.1. Una variable Y tiene distribución de Poisson con parámetro $\mu \geq 0$.

(a) Encuentre la función de verosimilitud correspondiente a la observación $Y = y$.

(b) Determine la estimación de verosimilitud máxima correspondiente al dato $Y = y$.

(c) ¿Cuál es el estimador de verosimilitud máxima de μ?

Solución. (a) Dada la observación $Y = y$, la función de verosimilitud es

$$\ell(\mu|y) = f(y|\mu) = e^{-\mu}\frac{\mu^y}{y!}, \quad \mu \geq 0. \tag{4.1}$$

(b) La estimación de verosimilitud máxima de μ correspondiente al dato $Y = y$ es un maximizador de la función $\mu \mapsto \ell(\mu|y)$. Para encontrarlo, recuerde que los posibles valores de Y son enteros no negativos, y considere los dos siguientes casos exahustivos:

Caso 1: $y = 0$. En estas circunstancias $\ell(\mu|y) = \ell(\mu|0) = e^{-\mu}$ es una función estrictamente decreciente en el intervalo $[0, \infty)$, y por lo tanto se maximiza en el valor $\mu = 0$, esto es,

$$\hat{\mu}(0) = 0$$

Caso 2: $y = 1, 2, 3, \ldots$. En este contexto, la función de verosimilitud $\mu \mapsto \ell(\mu|y)$ en (4.1) se anula en $\mu = 0$ y además, $\lim_{\mu\to\infty} \ell(\mu|y) = 0$, propiedades que indican que la función $\ell(\cdot|y)$ asume su valor máximo en un punto del intervalo abierto $(0, \infty)$, y por lo tanto el maximizador debe satisfacer la ecuación de verosimilitud

$$\partial_\mu \ell(\mu|y) = 0.$$

la cual equivale a

$$-e^{-\mu}\frac{\mu^y}{y!} + e^{-\mu} y \frac{\mu^{y-1}}{y!} = 0;$$

vea (4.1). Cancelando el término no nulo $e^{-\mu}/y!$ en ambos lados de esta igualdad, se obtiene $\mu^y - y\mu^{y-1} = 0$, esto es

$$\mu^{y-1}(-\mu + y) = 0,$$

y debido a que la solución que se busca pertenece al intervalo $(0, \infty)$, y por lo tanto es no nula, se desprende que $-\mu + y = 0$, es decir, $\mu = y$. En resumen, se ha mostrado que la ecuación $\partial_\mu \ell(\mu|y) = 0$ tiene una única solución en el intervalo $(0, \infty)$ la cual está dada por $\mu = y$. Combinando este hecho con la propiedad de que $\ell(\cdot|y)$ alcanza su máximo en un punto de $(0, \infty)$, se desprende que el maximizador de $\ell(\cdot|y)$ es $\mu = y$. Por lo tanto,

$$\hat{\mu}(y) = y, \quad y = 1, 2, 3, \ldots$$

Combinando esta especificación con la obtenida en el caso anterior, se desprende que, dada la observación $Y = y$, la estimación de verosimilitud máxima es $\hat{\mu}(y) = y$ para cualquier entero no negativo y.

(c) Recuerde que el estimador de verosimilitud máxima $\hat{\mu}(Y)$ se obtiene sustituyendo la variable aleatoria Y por la observación y en la fórmula para la estimación $\hat{\mu}(y)$. Puesto que $\hat{\mu}(y) = y$, se tiene que $\hat{\mu}(Y) = Y$. □

Ejemplo 4.2. Al utilizar el Cálculo para obtener una estimación o estimador de verosimilitud máxima, frecuentemente es más simple maximizar el logaritmo L de la función de verosimilitud ℓ en vez de optimizar ℓ directamente:

$$L = \log(\ell).$$

Ambas funciones toman su valor máximo en los mismos puntos, así que es suficiente maximizar cualquiera de ellas.

(a) Calcule $L(\pi)$ para la distribución binomial.

(b) Cuando el maximizador de L se encuentra en un punto interior, entonces el maximizador puede encontrarse resolviendo *la ecuación de verosimilitud*, la cual es la ecuación que resulta tomado la derivada parcial de L respecto al parámetro, e igualando esta derivada a cero. Determine la ecuación de verosimilitud para la distribución binomial, y resuélvala para mostrar que la estimación de verosimilitud máxima de π es $\hat{\pi} = y/N$.

Solución. (a) La función de probabilidad binomial con parámetros π y N es

$$f(y;\pi) = \binom{N}{y}\pi^y(1-\pi)^{N-y}, \quad y = 0, 1, 2, \ldots, N$$

donde el parámetro π se ubica entre 0 y 1. Luego, dada la observación $Y = y$, la función de verosimilitud es

$$\ell(\pi|y) = f(y;\pi) = \binom{N}{y}\pi^y(1-\pi)^{N-y}, \quad \pi \in [0, 1]$$

y el correspondiente logaritmo es

$$L(\pi|y) = \log\binom{N}{y} + y\log(\pi) + (N-y)\log(1-\pi)$$

(b) La ecuación de verosimilitud se obtiene calculando $\partial_\pi L(\pi|x)$ e igualando esta derivada a cero. Note que $\partial_\pi L = \dfrac{y}{\pi} - \dfrac{N-y}{1-\pi}$, de tal suerte que la ecuación de verosimilitud es

$$\frac{y}{\pi} - \frac{N-y}{1-\pi} = 0.$$

Multiplicando ambos lados de esta igualdad por $\pi(1-\pi)$ se obtiene $(1-\pi)y - \pi(N-y) = 0$; a partir de esta ecuación se desprende que $y - \pi y - N\pi + \pi y = 0$,

esto es $y - N\pi = 0$, cuya solución es $\pi = y/N$. Esta solución pertenece al interior del espacio de parámetros $(0, 1)$ cuando $y = 1, 2, \ldots, N-1$, de donde se desprende que

$$\hat{\pi} = \frac{y}{N}, \quad \text{si } y = 1, 2, \ldots, N-1.$$

Pero ¿qué sucede cuando $y = 0$, o $y = N$, los otros valores posibles de Y? La estimación de verosimilitud máxima debe calcularse separadamente en cada caso. De acuerdo a la parte (a),

$$L(\pi|0) = \log \binom{N}{0} + 0 \log(\pi) + (N - 0) \log(1 - \pi) = N \log(1 - \pi),$$

función que es continua en $\pi \in [0, 1)$, y decreciente en ese intervalo, pues

$$D_\pi L(\pi|0) = -\frac{N}{1 - \pi}, \quad \text{para } \pi \in (0, 1).$$

Luego, $L(\pi|0)$ se maximiza en $\pi = 0$, lo cual significa que

$$\hat{\pi}(0) = 0 = 0/N, \quad \text{si } y = 0.$$

Por otro lado,

$$L(\pi|N) = \log \binom{N}{N} + N \log(\pi) + (N - N) \log(1 - \pi) = N \log(\pi)$$

es una función continua y creciente en $\pi \in [0, 1]$, pues $D_\pi L(\pi|N) = N/\pi$ para $\pi \in (0, 1)$, de donde se desprende que su maximizador es $\pi = 1$, de manera que

$$\hat{\pi}(N) = 1 = N/N.$$

A partir de la expresión de $\hat{\pi}$ en cada caso, se desprende que

$$\hat{\pi}(y) = y/N, \quad y = 0, 1, \ldots, N,$$

y entonces el estimador de verosimilitud máxima, obtenido sustituyendo el dato y por la variable aleatoria Y en la fórmula de $\hat{\pi}(y)$, es $\hat{\pi}(Y) = Y/N$. □

5. Intervalo de Confianza Alternativo en el Caso Binomial

El problema de estimar proporciones es, no obstante su sencillez, muy importante en las aplicaciones reales. En esta sección se aborda el problema de construir un intervalo de confianza para una proporción mediante una variante del procedimiento delineado en los ejemplos anteriores (Brown *et. al.*, 2001).

Ejemplo 5.1. Muestre que los valores de π_0 para los cuales el valor absoluto del estadístico

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/N}}$$

toma un valor específico z_0 son soluciones de la ecuación

$$(1 + z_0^2/N)\pi_0^2 - (z_0^2/N + 2\hat{\pi})\pi_o + \hat{\pi}^2 = 0, \tag{5.1}$$

la cual es una ecuación cuadrática en π_0. Las soluciones de esta ecuación, denotadas por π_{0+} y π_{0-}, donde $\pi_{0-} \leq \pi_{0+}$, permiten obtener un intervalo de $100 \times (1-\alpha)\%$ de confianza para π de acuerdo al siguiente procedimento:

(1) Ponga $z_0 = z_{\alpha/2}$, el percentil bilateral de orden α de la distribución normal estándar.

(1) Escriba la ecuación (5.1) con los datos específicos.

(3) Resuelva la ecuación resultante para determinar las soluciones π_{0-} y π_{0+}.

(4) El intervalo de confianza deseado es

$$\pi_{0-} \leq \hat{\pi} \leq \pi_{0+}.$$

Utilice este procedimiento para encontrar un intervalo de 95% de confianza para π con los datos del Ejemplo 3.4.

Solución. Observe las siguientes equivalencias:

$$\begin{aligned}
|z| = z_0 &\iff |z|^2 = z_0^2 \\
&\iff |z|^2 = z_0^2 \\
&\iff \frac{(\hat{\pi} - \pi_0)^2}{\pi_0(1-\pi_0)/N} = z_0^2 \\
&\iff \hat{\pi}^2 - 2\hat{\pi}\pi_0 + \pi_0^2 = \pi_0 z_0^2/N - \pi_0^2 z_0^2/N \\
&\iff \hat{\pi}^2 - 2\hat{\pi}\pi_0 + \pi_0^2 = \pi_0 z_0^2/N - \pi_0^2 z_0^2/N
\end{aligned}$$

y después de una transposición se llega a

$$(1 + z_0^2/N)\pi_0^2 - (z_0^2/N + 2\hat{\pi})\pi_0 + \hat{\pi}^2 = 0.$$

Ahora se usará el procedimiento delineado en el enunciado del ejercicio para determinar un intervalo de confianza para π_0 con nivel aproximado de 95%:

(1) Se pone $z_0 = 1.96 = z_{0.05/2}$;

(2) Se plantea la ecuación cuadrática usando los datos específicos: $N = 976$ y $\hat{\pi} = 0.4027$. Con estos datos se tiene

$$a = (1 + z_0^2/N) = 1.003936066, \quad b = -(z_0^2/N + 2\hat{\pi}) = -0.809336066$$

y

$$c = \hat{\pi}^2 = 0.16216729$$

de manera que es necesario resolver la ecuación

$$1.003936066\pi_0^2 - 0.809336066\pi_0 + 0.16216729 = 0.$$

(3) Utilizando la fórmula cuadrática se obtienen las soluciones

$$\pi_{0-} = 0.372370141, \quad \pi_{0+} = 0.433792814$$

Por lo tanto,

(4) El intervalo de confianza para π_0, con nivel aproximado de 95%, es

$$0.3724 \leq \pi_0 \leq 0.4338,$$

el cual es muy parecido al obtenido en el Ejemplo 3.4 por otro método. El procedimiento delineado en este ejemplo, produce intervalos de confianza cuyo nivel real de confianza se acerca más al nominal $100 \times (1 - \alpha)\%$ que el obtenido por el método del Ejemplo 3.4. Ambas técnicas producen intervalos que se asemejan más y más conforme el tamaño de la muestra crece. En el caso presente ambos métodos producen intervalos 'similares', pues el tamaño de la muestra $N = 976$ es 'grande'. □

Capítulo 3

Medidas de Asociación

Este capítulo trata sobre un tema fundamental en el análisis estadístico, a saber, determinar si existe o no asociación entre dos variables. En otras palabras, se trata de determinar si la distribución de una variable aleatoria Y se altera cuando cambia el valor de otra variable X. Las medidas que se analizan son tres: La primera de ellas es la diferencia de probabilidades, luego, se estudia la denominada tasa de riesgo, la cual se obtiene como el cociente de dos probabilidades y, finalmente, se considera la razón de oportunidades, la cual desempeña un importante papel en el estudio de los denominados modelos log-lineales (Agresti, 1996, 2004, Dobson, 2001). La exposición concluye con la prueba ji-cuadrada de independencia de dos variables categóricas.

1. Clasificación Doble

Suponga que se realiza un experimento aleatorio, como resultado del cual se obtienen dos variables categóricas X y Y cuyas posibles categorías son $x_1, x_2, \ldots, x_r$ y $y_1, y_2, \ldots, y_c$, respectivamente. Puede pensarse que se selecciona un objeto, el cual es clasificado de acuerdo a su color (X) y a su dureza (Y). De esta manera, cada objeto seleccionado es clasificado de acuerdo a dos patrones, y el resultado final será el registro de las dos clasificaciones, por ejemplo, (verde, blando), o (azul, medianamente duro). El conjunto de todas las parejas (x_i, y_j) define una partición del espacio muestral, y es posible definir una nueva variable categórica Z mediante

$$Z = (x_i, y_j) \iff X = x_i, \quad Y = y_j.$$

De esta forma, Z engloba a las dos variables aleatorias originales X y Y. Los

diversos valores posibles de Z pueden visualizarse en una tabla cuyas lineas se etiquetan mediante $x_1, \ldots, x_r$ y cuyas columnas se identifican usando $y_1, \ldots, y_c$:

	y_1	y_2	$\cdots$	y_c
x_1	$*$	$*$	$\cdots$	$*$
x_2	$*$	$*$	$\cdots$	$*$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_r	$*$	$*$	$\cdots$	$*$

Valores Posibles de Z

En adelante, π_{ij} denota la probabilidad de que Z tome el valor (x_i, y_j), esto es,

$$\pi_{ij} = P[Z = (x_i, y_j)]; \tag{1.1}$$

esta función de probabilidad usualmente se describe mediante la siguiente tabla:

	y_1	y_2	$\cdots$	y_c
x_1	π_{11}	π_{12}	$\cdots$	π_{1c}
x_2	π_{21}	π_{22}	$\cdots$	π_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_r	π_{m1}	π_{m2}	$\cdots$	π_{rc}

(1.2)

Función de Probabilidad de Z

Para estimar las probabilidades π_{ij} se toma una muestra de tamaño N de la población, obteniendo los valores $Z_1, Z_2, \ldots, Z_N$ en cada repetición. Como se discutió en el Capítulo 1, para estimar las probabilidades π_{ij} es suficiente con registrar la frecuencia n_{ij} con que se observa cada uno de los posibles valores (x_i, y_j) de Z, las cuales naturalmente se acomodan en forma tabular de la siguiente forma, dando lugar a lo que se conoce como *tabla de contingencia*, término acuñado por el célebre estadístico Karl Pearson; vea Pearson (1922).

	y_1	y_2	$\cdots$	y_c
x_1	n_{11}	n_{12}	$\cdots$	n_{1c}
x_2	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

(1.3)

Tabla de Contingencia Para Muestras de Z

De conformidad a los resultados del Capítulo 1, el estimador de verosimilitud máxima de π_{ij} es

$$\hat{\pi}_{ij} = \frac{n_{ij}}{N}, \tag{1.4}$$

donde $N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$.

2. Distribución Condicional

Dado un nivel de X, digamos x_i, la distribuición condicional de Y está dada por

$$\begin{aligned} \pi_{j|i} &= P[Y = y_j | X = x_i] \\ &= \frac{P[X = x_i, Y = y_j]}{P[X = x_i]} = \frac{\pi_{ij}}{\pi_{i+}}, \quad j = 1, 2, \ldots, c. \end{aligned} \tag{2.1}$$

donde

$$\pi_{i+} = \sum_{j=1}^{c} \pi_{ij},$$

de manera que el símbolo de adición indica suma sobre el índice que sustituye. Note que el estimador de verosimilitud máxima de $\pi_{j|i}$ es

$$\hat{\pi}_{j|i} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}} \tag{2.2}$$

donde

$$\hat{\pi}_{i+} = \sum_{j=1}^{c} \hat{\pi}_{i\,j} = \frac{1}{N} \sum_{j=1}^{c} n_{i\,j} = \frac{n_{i+}}{N}.$$

Se dice que las variables X y Y están asociadas, si la distribución condicional de Y varía cuando cambia el nivel x_i de X. Por el contrario, si, la distribución condicional de Y no depende del valor que X toma, se dice que X y Y son independientes. Un problema fundamental en el análisis de datos es determinar si X y Y están asociadas y, naturalmente, una medida de asociación debe considerar las diferencias entre las probabilidades condicionales $\pi_{j|i}$ y $\pi_{j|i_1}$ del evento $[Y = y_j]$ bajo diferentes niveles x_i y x_{i_1} de X. Tres de dichas medidas se presentan en la siguiente sección.

3. Medidas de Asociación

En el desarrollo subsecuente, los índices i, i_1 son enteros distintos entre 1 y r, mientras que j denota un índice arbitrario entre 1 y c. Con esta notación, se definirán tres medidas de asociación entre las variables X y Y; detalles sobre estas medias pueden encontrarse en Agresti (1996, 2004), o en Bishop *et. al.*, (1975).

1. Diferencia de Probabilidades. La diferencia de probabilidades se denota mediante δ y se define como

$$\delta = \pi_{j|i} - \pi_{j|i_1}.$$

Su estimador de verosimilitud máxima es

$$\hat{\delta} = \hat{\pi}_{j|i} - \hat{\pi}_{j|i_1},$$

el cual tiene error estándar asintótico dado por

$$\text{ASE}(\hat{\delta}) = \sqrt{\frac{\hat{\pi}_{j|i}(1-\hat{\pi}_{j|i})}{n_{i+}} + \frac{\hat{\pi}_{j|i_1}(1-\hat{\pi}_{j|i_1})}{n_{i_1+}}},$$

de manera que

$$\sqrt{n_{i_1+} + n_{i+}}\ \frac{\hat{\delta} - \delta}{\text{ASE}(\hat{\delta})} \xrightarrow{\text{d}} \mathcal{N}(0,1)$$

2. Cociente de Probabilidades (Tasa de Riesgo). El cociente de probabilidades, también denominado razón de riesgo, se define como

$$\rho = \frac{\pi_{j|i}}{\pi_{j|i_1}}.$$

Su estimador de verosimilitud máxima es

$$\hat{\rho} = \frac{\hat{\pi}_{j|i}}{\hat{\pi}_{j|i_1}},$$

con error estándar asintótico

$$\text{ASE}(\hat{\rho}) = \sqrt{\frac{1-\hat{\pi}_{j|i}}{\hat{\pi}_{j|i}\ n_{i+}} + \frac{1-\hat{\pi}_{j|i_1}}{\hat{\pi}_{j|i_1}\ n_{i_1+}}},$$

lo cual significa que

$$\sqrt{n_{i_1+}+n_{i+}}\ \frac{\hat{\rho}-\rho}{\mathrm{ASE}(\hat{\rho})}\xrightarrow{\mathrm{d}}\mathcal{N}(0,1).$$

3. Razón de Oportunidades. La razón de oportunidades se denota por θ y se define como

$$\theta=\frac{\pi_{j|i}/(1-\pi_{j|i})}{\pi_{j|i_1}/(1-\pi_{j|i_1})}.$$

Su estimador de verosimilitud máxima es

$$\hat{\theta}=\frac{\hat{\pi}_{j|i}/(1-\hat{\pi}_{j|i})}{\hat{\pi}_{j|i_1}/(1-\hat{\pi}_{j|i_1})}.$$

con error estándar asintótico

$$\mathrm{ASE}(\hat{\theta})=\sqrt{\frac{1}{n_{ij}}+\frac{1}{n_{i+}-n_{ij}}+\frac{1}{n_{i_1\,j}}+\frac{1}{n_{i_1+}-n_{i_1\,j}}},$$

de tal manera que

$$\sqrt{n_{i_1+}+n_{i+}}\ \frac{\hat{\theta}-\theta}{\mathrm{ASE}(\hat{\theta})}\xrightarrow{\mathrm{d}}\mathcal{N}(0,1).$$

4. Prueba de Independencia

Probar la hipótesis de que X y Y no están asociadas, esto es, que son independientes, es de la mayor importancia. Esta hipótesis se expresa de la siguiente forma:

$$\mathcal{H}\colon \pi_{ij}=\pi_{i+}\pi_{+j}\quad \text{para todo } i,\ j.$$

Se presentarán dos estadísticos para probar esta hipótesis (Goodman y Kruskal, 1979, Agresti, 1996, 2004). El procedimiento de cálculo de ambos presenta las mismas dos fases iniciales y, como punto de partida es conveniente notar que, cuando $\mathcal{H}$ es cierta, la frecuencia esperada de la observación (x_i, y_j) es $\mu_{i\,j}=N\pi_{i+}\pi_{+j}$ el cual tiene como estimador de verosimilitud máxima al estadístico

$$\hat{\mu}_{i\,j}=N\hat{\pi}_{i+}\hat{\pi}_{+j}.$$

Primero, se calculan las estimaciones $\hat{\mu}_{i\,j}$ y se colocan en la siguiente tabla de frecuencias esperadas bajo independencia:

$$\begin{array}{cccccc} & y_1 & y_2 & \cdots & y_c \\ x_1 & \hat{\mu}_{11} & \hat{\mu}_{12} & \cdots & \hat{\mu}_{1c} \\ x_2 & \hat{\mu}_{21} & \hat{\mu}_{22} & \cdots & \hat{\mu}_{2c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_r & \hat{\mu}_{r1} & \hat{\mu}_{r2} & \cdots & \hat{\mu}_{rc} \end{array} \tag{4.1}$$

Tabla de Frecuencias Esperadas Bajo Independencia

A continuación, estas frecuencias esperadas se comparan con las frecuencias observadas de la tabla original, la cual se reproduce a continuación:

$$\begin{array}{cccccc} & y_1 & y_2 & \cdots & y_c \\ x_1 & n_{11} & n_{12} & \cdots & n_{1c} \\ x_2 & n_{21} & n_{22} & \cdots & n_{2c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_r & n_{r1} & n_{r2} & \cdots & n_{rc} \end{array} \tag{4.2}$$

Tabla de Frecuencias Observadas

Estadístico X^2. Se recorren las posiciones de la tabla calculando la diferencia entre lo observado y lo esperado, cantidad que se eleva al cuadrado y se divide por la frecuencia esperada para obtener

$$X^2 = \sum_{i\,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

Bajo $\mathcal{H}$, X^2 tiene una distribución ji-cuadrada con $(r-1)(c-1)$ grados de libertad, y la hipótesis $\mathcal{H}$ se rechaza al nivel de significancia α si y sólo si

$$X^2 > \chi^2_{(r-1)(c-1),\,\alpha}.$$

Estadístico G^2. Se recorren las posiciones de la tabla calculando el cociente entre lo esperado y lo observado tomando el logaritmo natural de esta cantidad, la cual se multiplica por el doble de la frecuencia observada para obtener

$$G^2 = 2\sum_{i\,j} n_{i\,j} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$$

Como en el caso anterior, bajo $\mathcal{H}$ este estadístico tiene la distribución ji-cuadrada con $(r-1)(c-1)$ grados de libertad, y la hipótesis $\mathcal{H}$ se rechaza al nivel de significancia α si y sólo si

$$G^2 > \chi^2_{(r-1)(c-1),\,\alpha}.$$

Los dos procedimientos delineados para probar el supuesto de independencia $\mathcal{H}$ tienen un nivel de significancia aproximadamente igual a α, y la aproximación es mejor conforme el número de datos N crece. Además, bajo el supuesto $\mathcal{H}$, la diferencia entre G^2 y X^2 se aproxima a cero a medida que N se incrementa sin límite, razón por la cual ambos métodos de prueba conducen a conclusiones similares de aceptación o rechazo de $\mathcal{H}$ (Agresti, 1996, 2004).

Capítulo 4

Ejemplos del Análisis de Tablas de Contingencia

En este capítulo se discuten de forma detallada ejemplos de la aplicación de las ideas presentadas en el capítulo precedente para estudiar la asociación entre dos variables categóricas.

1. Las Medidas de Asociación

El siguiente ejemplo se refiere al uso de la aspirina como un atenuante del riesgo de sufir un ataque al corazón, e ilustra la diferencia de probabilidades, las razones de riesgo y de oportunidades como medidas de asociación.

Ejemplo 1.1. Un estudio realizado en Suecia consideró el efecto de una dosis baja de aspirina en la reducción de ataques al corazón entre gente que ha padecido enfermedades cardiacas. De 1360 pacientes, 676 fueron asignados aleatoriamente al tratamiento de aspirina (una tableta diaria de contenido bajo) y 684 a un tratamiento de placebo. Durante un período de seguimiento de tres años, el número de muertes debidas a infarto al miocardio fue de 18 para el grupo de aspirina y de 28 en el grupo de placebo.

(a) Calcule la diferencia de proporciones, riesgo relativo de muerte y la razón de oportunidades e interprete los resultados.

(b) Realice un análisis inferencial para estos datos e interprete los resultados.

Solución. Las cifras de este estudio pueden resumirse en la siguiente tabla de contingencia:

	Muerte	Sobrevivencia	Total
1: Aspirina	18	658	676
2: Placebo	28	684	684

(a) Las proporciones de muerte en los grupos de aspirina y placebo, denotadas por p_1 y p_2, respectivamente, son

$$p_1 = 18/676 = 0.0266, \qquad p_2 = 28/684 = 0.0409.$$

La diferencia de proporciones es

$$p_1 - p_2 = 0.0266 - 0.0409 = -0.0143;$$

esta cifra indica que, *en la muestra analizada*, la proporción de muertes dentro del grupo de aspirina fue menor que la proporción correspondiente en el grupo del placebo.

La tasa de riesgo muestral es

$$r = p_1/p_2 = 0.6505,$$

la cual indica que la proporción de muertes en el grupo de aspirina es sólo el 65% de la proporción de muerte en el grupo del placebo, esto es, *dentro de la muestra*, en el grupo de aspirina se observa una disminución de 35% en la tasa de muerte respecto a la proporción de muerte en el grupo del placebo.

La razón de oportunidades en la muestra es

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{0.0266/(1-0.0266)}{0.0409/(1-0.0409)} = 0.6409,$$

de manera que la razón de oportunidades de muerte en el grupo de aspirina fue de sólo el 64% de la razón de muerte en el grupo del placebo, es decir, dentro de la muestra analizada, la ingesta de aspirina disminuye la razón de oportunidades de muerte en un 36%.

De acuerdo a estas cifras, dentro de la muestra la ingestión de aspirina tuvo efectos beneficos en el sentido de que la mortalidad fue menor que en el grupo del placebo. Para extender estas conclusiones a la población de interés, es necesario realizar el análisis inferencial del siguiente apartado, el cual consiste en establecer intervalos de confianza para cada una de las cantidades consideradas en este inciso.

(b) Denote mediante π_1 y π_2 a las probabilidades de muerte dentro de los grupos de aspirina y placebo, respectivamente.

(i) Inferencia para la diferencia de probabilidades $\delta = \pi_1 - \pi_2$. El intervalo de confianza correspondiente es

$$\hat{\delta} - z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \leq \delta \leq \hat{\delta} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

donde $\hat{\pi}_i = p_i$ es el estimador de π_i, $i = 1, 2$, $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_2$ es el estimador de δ, y $z_{\alpha/2}$ es el percentil bilateral de orden α en la distribución normal estándar, y $1 - \alpha$ es el nivel de confianza deseado. En este ejemplo se tomará $\alpha = 0.05$, lo que corresponde a un nivel de confianza de $1 - 0.05 = 95\%$, de manera que

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

Note además que $n_1 = 676$ y $n_2 = 684$ son los tamaños de muestra de los grupos de aspirina y placebo, respectivamente. Haciendo los cálculos se obtienen las siguientes cifras:

$$\hat{\delta} = p_1 - p_2 = 0.0266 - 0.0409 = -0.0143$$

$$\begin{aligned}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} &= \sqrt{\frac{0.0266(1-0.0266)}{676} + \frac{0.0409(1-0.0409)}{684}} \\ &= 0.00978\end{aligned}$$

y por lo tanto

$$z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} = 1.96(0.00978) = 0.0192$$

Luego, el intervalo de 95% de confianza para δ es $-0.0143 - 0.0192 \leq \delta \leq -0.0143 + 0.0192$, esto es,

$$-.0335 \leq \delta \leq .0049$$

El analista tiene confianza en que el intervalo $[-0.0143, .0049]$ contiene al verdadero valor de la diferenca δ; note que dicho intervalo contiene valores positivos, de manera que no puede concluirse que δ sea negativo, como lo suguiere la discusión del inciso precedente, lo cual ilustra el hecho de que el análisis estadístico riguroso debe realizarse antes de establecer conclusiones para la poblacón de interés a partir de estimadores puntuales.

(ii) Inferencia para la tasa de oportunidades θ. El intervalo de confianza para la tasa de oportuidades poblacional θ se obtiene a partir de

$$\log(\hat{\theta})) - z_{\alpha/2}\text{ASE}(\log(\hat{\theta})) \leq \log(\theta) \leq \log(\hat{\theta})) + z_{\alpha/2}\text{ASE}(\log(\hat{\theta}))$$

donde $\hat{\theta} = 0.6409$ es la estimación puntual obtenida previamente,

$$\text{ASE}(\log(\hat{\theta})) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

y los $n_{i\,j}$ son las frecuencias en la tabla de contingencia. En este caso,

$$\text{ASE}(\log(\hat{\theta})) = \sqrt{\frac{1}{18} + \frac{1}{658} + \frac{1}{28} + \frac{1}{656}} = 0.3071.$$

Puesto que $\log(\hat{\theta}) = \log(0.6409) = -0.4449$, usando una confianza de 95%, de manera que $z_{\alpha/2} = 1.96$, los extremos para el intervalo de confianza para $\log(\theta)$ son

$$-0.4449 - 1.96(.3071) = -1.0468 \quad \text{y} \quad -0.4449 + 1.96(.3071) = 0.1570.$$

de donde se desprende que el analista tiene una confianza de 95% en que

$$-1.0468 \leq \log(\theta) \leq 0.1570$$

y por lo tanto, tomando exponencial en cada término se llega a $e^{-1.0468} \leq \theta \leq e^{0.1570}$, esto es,

$$0.3511 \leq \theta \leq 1.1700$$

Este es el intervalo de 95% de confianza para θ; note que este intervalo contiene puntos mayores que 1, de manera que no puede concluirse que la tasa de oportunidades de morir de infarto sean menores en la población que toma aspirina que en la población que no la ingiere; sin embargo, note que 'la mayor parte' del intervalo consiste de puntos menores que uno.

(iii) Inferencia para la tasa de riesgo $\rho = \pi_1/\pi_2$. El intervalo de confianza para la tasa de riesgo ρ se obtiene a partir de

$$\log(\hat{\rho})) - z_{\alpha/2}\text{ASE}(\log(\hat{\rho})) \leq \log(\rho) \leq \log(\hat{\rho})) + z_{\alpha/2}\text{ASE}(\log(\hat{\rho}))$$

donde $\hat{\rho} = \hat{\pi}_1/\hat{\pi}_2$ y

$$\text{ASE}(\log(\hat{\rho})) = \sqrt{\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}}$$

donde $\hat{\pi}_1 = .0266$ y $\hat{\pi}_1 = .0409$ son las estimaciones de π y π_2, respectivamente, mientras que $n_1 = 676$ y $n_2 = 684$ son los tamaños de muestra de los grupos 1 y 2. En el caso presente estos datos dan el siguente resultado:

$$\text{ASE}(\log(\hat{\rho})) = \sqrt{\frac{1-0.0266}{676(.0266)} + \frac{1-0.0409}{684(0.0409)}} = 0.29734$$

y por lo tanto, usando $z_{\alpha/2} = 1.96$, lo que equivale a estipular una confianza de 95%, se tiene que

$$1.96\text{ASE}(\log(\hat{\rho})) = (1.96)0.29734 = 0.5828.$$

Puesto que $\hat{\rho} = \hat{\pi}_1/\hat{\pi}_2 = 0.6504$ se obtiene $\log(\hat{\rho}) = -0.4302$, y el intervalo de confianza con nivel de 95% para $\log(\rho)$ es $-0.4302 - 0.5828 \leq \log(\rho) \leq -0.4302 + 0.5828$, esto es,

$$-1.013 \leq \log(\rho) \leq 0.1526$$

y por lo tanto

$$0.3631 = e^{-1.013} \leq \rho = \frac{\pi_1}{\pi_2} \leq e^{0.1526} = 1.1649$$

Note que este intervalo contiene al 1, de manera que no es posible declarar que ρ, la tasa de riesgo de muerte sea menor a 1, esto es, que la ingesta de aspirina diminuya la probabilidad de muerte por infarto. □

2. Intepretación de las Razones de Oportunidades y de Riesgo

En el siguiente ejemplo se ilustra la diferencia entre las interpretaciones de tasa de riesgo y razón de oportunidades.

Ejemplo 2.1. La razón de oportunidades entre tratamiento (con niveles A y B) y respuesta (muerte, sobrevivencia) es igual a 2.0.

(a) Explique que hay de incorrecto en la siguiente interpretación: 'La probabilidad de muerte con el tratamiento A es el doble que con el tratamiento B'. Proporcione la interpretación correcta.

(b) ¿Cuándo es la interpretación entrecomillada en la parte (a) correcta en un sentido aproximado?

(c) Las oportunidades de muerte son de 0.5 para el tratamiento A. ¿Cuál es la probabilidad de muerte para (i) el tratamiento A, y (ii) para el tratamiento B?

Solución. Para analizar la respuesta, es conveniente introducir la siguiente tabla:

	Muerte	Sobrevivencia
Tratamiento A	π_A	$1-\pi_A$
Tratamiento B	π_B	$1-\pi_B$

donde π_A y π_B son las probabilidades de muerte bajo los tratamientos A y B, respectivamente. Las oportunidades de muerte en los tratamientos A y B son $\pi_A/(1-\pi_A)$ y $\pi_B/(1-\pi_B)$, respectivamente, y la correspondiente razón de oportunidades es

$$\theta = \frac{\pi_A/(1-\pi_A)}{\pi_B/(1-\pi_B)},$$

mientras que la tasa de riesgo de muerte es

$$\rho = \frac{\pi_A}{\pi_B}.$$

(a) Una tasa de oportunidades igual a 2, esto es, $\theta = 2$, significa que

$$\pi_A/(1-\pi_A) = 2\pi_B/(1-\pi_B), \tag{2.1}$$

es decir, *las oprtunidades de muerte* bajo el tratamiento A son el doble que *las oportunidades de muerte* bajo el tratamiento B. Por otro lado, $\rho = 2$ significa que $\pi_A = 2\pi_B$, de manera que la interpretación de una tasa de riesgo igual a 2 si es que la *probabilidad de muerte* bajo A es el doble que bajo B. Por lo tanto, el enunciado entre comillas corresponde al significado de $\rho = 2$ y no al de $\theta = 2$.

(b) Note que $\theta = 2$ equivale a

$$\frac{\pi_A}{\pi_B} = 2\frac{1-\pi_A}{1-\pi_B}$$

de manera que si π_A y π_B son pequeños, entonces $\pi_A/\pi_B \approx 2$, esto es, $\pi_A \approx 2\pi_B$, y en este caso, si es cierto que la probabilidad de muerte bajo A es, aproximadamente, el doble de la probabilidad de muerte bajo B.

(c) Suponga que $\theta = 2$ y que las oportunidades de muerte bajo el tratamiento A son iguales a 0.5. Esta última condición significa que $\pi_A/(1-\pi_A) = 0.5$, de donde se desprende que $\pi_A = 0.5(1-\pi_A)$, esto es, $1.5\pi_A = 0.5$, y entonces

$$\pi_A = 0.5/1.5 = 1/3.$$

Por otro lado, como $\theta = 2$ significa que (2.1) ocurre, la condición $\pi_A/(1-\pi_A) = 0.5$ implica que $0.25 = \pi_B/(1-\pi_B)$, es decir, $0.25(1-\pi_B) = \pi_B$ de donde se desprende que $0.25 = 1.25\pi_B$, y por lo tanto

$$\pi_B = \frac{0.25}{1.25} = 0.20.$$

Como comprobación, note que con estos resultados se tiene que la tasa de oportunidades de muerte bajo A y B son $\phi_A = \pi_A/(1-\pi_A) = (1/3)/(1-1/3) = (1/3)/(2/3) = 1/2 = 0.5$ y $\phi_B = (0.20)/(1-0.20) = 0.20/0.80 = .25$, de manera que la razón de oportunidades es $\theta = \phi_A/\phi_B = 0.5/0.25 = 2$, como se estipuló desde el principio. □

3. Prueba de Independencia

En los siguientes dos ejemplos se utilizan las pruebas X^2 y G^2 para probar el supuesto de independencia, y el análisis se complementa construyendo intervalos de confianza para medidas de asociación.

Ejemplo 3.1. La siguiente tabla fue tomada de la Encuesta Social General de 1991 realizada en los Estados Unidos, donde se registro para cada persona entrevistada su raza (Blanco, Negro) y su creencia en la vida después de la muerte (Si, No).

	Si	No
Blanco	621	239
Negro	89	42

(a) Identifique cada clasificación como una variable de respuesta o como una variable explicatoria.

(b) Utilice el estadístico X^2 y el estadístico G^2 para probar la hipótesis de que la creencia en la vida después de la muerte no depende de la raza.

(c) Describa la asociación entre la raza y la creencia en la vida posterior a la muerte mediante un estimador puntual. Interprete la dirección y la intensidad de la asociación.

(d) Obtenga un intervalo de 95% de confianza para una medida de asociación. Interprete los resultados.

Solución. (a) En este caso la raza es la variable explicatoria, y la creencia en la vida después de la muerte es la variable de respuesta.

(b) Para evaluar el estadístico X^2 o el estadístico G^2 es conveniente evaluar primero las frecuencias esperadas de cada celda bajo la hipótesis de independencia. Con este fin, se calculan los totales de fila y columna de la tabla original de contingencia, anotando en la esquina sureste de la tabla el gran total de los datos, el cual aparece en negritas en la siguiente tabla:

			Total de fila
	621	239	860
	89	42	131
Total de columna	710	281	**991**

Usando esta información sobre los totales, se construye la tabla de frecuencias esperadas poniendo en la fila i y la columna j de la tabla, el producto del total de la fila i por el total de la columna j entre el gran total:

Tabla de frecuencias Esperadas

616.1453078	243.8546922
93.85469223	37.14530777

Por ejemplo, 616.1453078 = 860(710)/991.

Cálculo del estadístico X^2: Para evaluar X^2, para cada posición de la tabla se calcula la diferencia entre la frecuencia observada y la frecuencia esperada, se eleva al cuadrado dicha diferencia y se divide por la frecuencia esperada: En este caso se obtiene la siguiente tabla:

Cálculo de los términos de X^2

0.038250777	0.09664787
0.25111197	0.634482201

Por ejemplo, $(621 - 616.1453078)^2/616.1453078 = 0.038250777$. Sumando cada uno de los términos en la tabla anterior se obtiene el estadístico X^2:

$$X^2 = 0.038250777 + 0.09664787 + 0.25111197 + 0.634482201 = 1.02$$

En la presente tabla de contingencia con 2 lineas y dos columnas, X^2 tiene, aproximadamente, distribución ji-cuadrada con un grado de libertad, y la probabilidad de observar un valor de X^2 mayor o igual a 1.02 es

$$p = P[X^2 \geq 1.02] = 0.31204$$

de manera que no hay evidencia de una asociación entre raza y creencia en la vida después de la muerte.

Cálculo del estadístico G^2: Para evaluar G^2, para cada posición de la tabla se calcula el logaritmo natural del cociente de la frecuencia observada entre la esperada y se multiplica por el doble de la frecuencia observada. En este caso se obtiene la siguiente tabla:

Cálculo de los términos de G^2

$$\begin{array}{rr} 9.74753517 & -9.61208876 \\ -9.453827273 & 10.31790145 \end{array}$$

Por ejemplo, $2(621)\log(621/616.1453078) = 9.74753517$. Sumando los términos en la tabla anterior se obtiene el estadístico G^2:

$$G^2 = 9.74753517 - 9.61208876 - 9.453827273 + 10.31790145 = 0.999.$$

De nueva cuenta, debido a que la presente tabla de contingencia consiste de dos lineas y dos columnas, G^2 tiene, aproximadamente, distribución ji-cuadrada con un grado de libertad, y la probabilidad de observar un valor de G^2 mayor o igual a 0.999 es

$$p = P[G^2 \geq 0.999] = 0.317423$$

de manera que no hay evidencia de una asociación entre raza y creencia en la vida después de la muerte.

(c) Se considerarán la diferencia de probabilidades , la tasa de probabilidades, y la razón de oportunidades. Denote mediante π_1 la probabilidad de creer en la vida

después de la muerte para una persona de raza blanca, y mediante π_2 la correspondiente probabilidad para personas de raza negra: Note que las estimaciones correspondientes obtenidas de los datos muestrales son

$$\hat{\pi}_1 = 621/860 = 0.722093023, \qquad \hat{\pi}_2 = 89/131 = 0.679389313$$

(i) Diferencia de probabilidades $\delta = \pi - \pi_2$: La estimación de δ es

$$\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_2 = 0.722093023 - 0.679389313 = 0.04270371;$$

aunque esta cifra es positiva, y por lo tanto, a primera vista se sospecha que las personas de raza blanca tienen mayor propensión a creer en la vida después de la muerte, antes de establecer generalizaciones acerca de toda la población, es necesario construir un intervalo de confianza para δ: Con este fin, se usará la fórmula para dicho intervalo con nivel de confianza de 95%:

$$\hat{\delta} - 1.96\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \leq \delta \leq \hat{\delta} + 1.96\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

donde n_1 y n_2 son los números de personas de raza blanca y negra en la muestra, respectivamente. Con los datos del problema se obtiene

$$1.96\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} = 0.085346393$$

de manera que el intervalo de 95% de confianza para δ es

$$0.04270371 - 0.085346393 \leq \delta \leq 0.04270371 + 0.085346393,$$

esto es, $-0.043 \leq \delta \leq 0.128$; puesto que este intervalo contiene al origen, no es posible declarar que δ tenga un signo definido, de manera que, al nivel de significancia de 5%, no es posible concluir que exista diferencia entre las probabilidades de aceptar la existencia de vida después de la muerte para personas de raza blanca o negra.

(ii) La tasa de probabilidades $\rho = \pi_1/\pi_2$: La estimación es

$$\hat{\rho} = \hat{\pi}_1/\hat{\pi}_2 = 0.722093023/0.679389313 = 1.06$$

mientras que un intervalo de confianza con nivel de 95% para $\log(\rho)$ está dado por

$$\log(\hat{\rho}) - 1.96\text{ASE}(\log(\hat{\rho})) \leq \log(\rho) \leq \log(\hat{\rho}) + 1.96\text{ASE}(\log(\hat{\rho}))$$

donde

$$\begin{aligned}\text{ASE}(\log(\hat{\rho})) &= \sqrt{\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}} \\ &= \sqrt{\frac{1-0.7221}{860(0.7221)} + \frac{1-0.6794}{131(0.6794)}} \\ &= 0.06364;\end{aligned}$$

puesto que $\log(\hat{\rho}) = 0.06096$ se llega a

$$-0.0638 = 0.06096 - 1.96(0.6364) \leq \log(\rho) \leq 0.06096 + 1.96(0.6364) = 0.1857$$

y tomando exponencial se obtiene, finalmente,

$$0.9382 = e^{-0.0638} \leq \rho \leq e^{0.1857} = 1.2041;$$

este es el intervalo de 95% confianza para ρ, el cual contiene al uno y, por lo tanto, no es posible establecer diferencia entre las probabilidades de creencia en la vida después de la muerte para las dos razas consideradas.

(iii) Inferencia para la razón de oportunidades θ. La estimación puntual de θ es

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{0.7221/(1-0.7221)}{0.6794/(1-0.6794)} = 1.2262$$

y el error estándar asintótico de $\log(\hat{\theta}))$ es

$$\begin{aligned}\text{ASE}(\log(\hat{\theta})) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + +\frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{621} + \frac{1}{239} + \frac{1}{89} + \frac{1}{42}} = 0.2021\end{aligned}$$

El intervalo de confianza para la tasa de oportunidades poblacional θ se obtiene a partir de

$$\log(\hat{\theta})) - z_{\alpha/2}\text{ASE}(\log(\hat{\theta})) \leq \log(\theta) \leq \log(\hat{\theta})) + z_{\alpha/2}\text{ASE}(\log(\hat{\theta}))$$

donde $\hat{\theta} = 1.2262$ es la estimación puntual obtenida previamente. Puesto que

$$\log(\hat{\theta}) = \log(1.2262) = 0.2039,$$

usando una confianza de 95%, de manera que $z_{\alpha/2} = 1.96$, el intervalo de confianza para $\log(\theta)$ es

$$0.2039 - 1.96(.2021) \leq \log(\theta) \leq 0.2039 + 1.96(.2021)$$

es decir,, $-0.1921 \leq \log(\theta) \leq 0.6000$ y por lo tanto, tomando exponencial en cada término se llega a $e^{-0.1921} \leq \theta \leq e^{0.6000}$, y finalmente

$$0.8251 \leq \theta \leq 1.82211$$

Este es el intervalo de 95% de confianza para θ; note que este intervalo contiene al 1, de manera que no puede concluirse que la tasa de oportunidades de creer en la vida después de la muerte difiera para los dos razas consideradas. □

Ejemplo 3.2. La siguiente tabla se compiló a partir de los registros del Departamento de Seguridad Vial. La tabla registra si en un accidente automovilístico hubo fallecimientos y si los ocupantes usaban o no cinturón de seguridad en ese momento.

	Daño a los ocupantes	
	Fatal	No fatal
1: No se usaba cinturón	1601	162527
2: Si se usaba cinturón	510	412368

(a) Calcule e interprete la razón muestral de oportunidades, el riego relativo y la diferencia de proporciones.

(b) Construya intervalos de confianza para los indicadores poblacionales correspondientes y establezca conclusiones entre la muerte en un accidente al conducir un automóvil y el uso del cinturón de seguridad.

Solución. Como punto de partida, es conveniente completar la tabla original incluyendo totales de fila y columna así como el gran total:

	Fatal	No fatal	Total de fila
1: No se usaba cinturón	1601	162527	163858
2: Si se usaba cinturón	510	412368	412878
Total de columna	2111	574625	**576736**

Denote mediante π_1 y π_2 las probabilidades de fallecer en un accidente cuando no se usa cinturón de seguridad y cuando se usa, respectivamente. Las estimaciones muestrales de esos parámetros son, de acuerdo a la tabla,

$$\hat{\pi}_1 = \frac{1602}{163858} = 0.009770655$$
$$\hat{\pi}_2 = \frac{510}{412878} = 0.001235232$$

(a) La razón estimada de riesgo es

$$\hat{\rho} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = 7.901$$

lo cual significa que la probabilidad de fallecimiento en un accidente cuando no se usa cinturón es 7.9 veces (casi 8) la correspondiente probabilidad usándolo. La razón estimada de oportunidades es

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = 7.978158687$$

de manera que, en la muestra, las oportunidades de muerte en un accidente se multiplicaron casi por 8 al no utilizar el cinturón de seguridad.

La diferencia estimada de probabilidades es

$$\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_2 = 0.008535423$$

la cual es positiva, indicando que la probabilidad de fallecer en un accidente es mayor cuando no se utiliza el cinturón de seguridad que cuando se utiliza. La diferencia, sin embargo, es de 9 milésimos, pudiendo pensarse que es insignificante. Esto ilustra, de nueva cuenta, el hecho de que al comparar probabilidades pequeñas, el uso de la tasa de riesgo o de la razón de oportunidades es más adecuado.

(b) Intervalos de 95% de confianza. A continuación se determinan los intervalos de confianza para cada una de las medidas de asociación consideradas en el precedente apartado.

(i) Intervalo de confianza para la tasa riesgo $\rho = \pi_1/\pi_2$. El procedimiento inicia planteando el intervalo de confianza para $\log(\rho)$, el cual está dado por

$$\log(\hat{\rho}) - 1.96\text{ASE}(\log(\hat{\rho})) \leq \log(\rho) \leq \log(\hat{\rho}) + 1.96\text{ASE}(\log(\hat{\rho}))$$

donde

$$\text{ASE}(\log(\hat{\rho})) = \sqrt{\frac{1-\hat{\pi}_1}{163858(\hat{\pi}_1)} + \frac{1-\hat{\pi}_2}{412878(\hat{\pi}_2)}} = 0.024869856$$

Puesto que $\log(\hat{\rho}) = \log(7.901) = 2.068124938$, se desprende que

$$2.068124938 - 1.96(0.024869856) \leq \log(\rho) \leq 2.068124938 + 1.96(0.024869856)$$

y finalmente

$$2.019380019 \leq \log(\rho) \leq 2.116869857.$$

Por lo tanto, un intervalo de confianza para ρ, obtenido tomando la función exponencial en cada término, está dado por

$$7.534 \leq \rho \leq 8.305,$$

lo cual muestra que no utilizar el cinturón de seguridad multiplica la probabilidad de muerte en un accidente de auto por al menos 7.5 veces.

(ii) Intervalo de confianza para $\theta = [\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]$, la razón de oportunidades. Como en el caso anterior, el procedimiento inicia estableciendo el intervalo de 95% de confianza para $\log(\theta)$, el cual está dado por

$$\log(\hat{\theta}) - 1.96\text{ASE}(\log(\hat{\theta})) \leq \log(\theta) \leq \log(\hat{\theta}) + 1.96\text{ASE}(\log(\hat{\theta}))$$

donde

$$\text{ASE}(\log(\hat{\theta})) = \sqrt{\frac{1}{1601} + \frac{1}{510} + \frac{1}{162257} + \frac{1}{412368}} = 0.05093115$$

Como $\log(\hat{\theta}) = 2.076707644$, se desprende que

$$2.076707644 - 1.96(0.05093115) \leq \log(\theta) \leq 2.076707644 + 1.96(0.05093115)$$

y por lo tanto

$$1.97688259 \leq \log(\theta) \leq 2.176532697$$

de tal manera que tomando la función exponencial en cada término, se obtiene

$$7.220199546 \leq \theta \leq 8.815686552$$

el cual es el intervalo de 95% de confianza para θ. A partir de esta expresión se desprende que, al no utilizar el cinturón de seguridad, las oportunidades de morir en un accidente se multiplican por, al menos, siete.

(iii) Intervalo de confianza para $\delta = \pi_1 - \pi_2$, la diferenca de probabilidades. En este caso, el estimador puntual de δ es

$$\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_2 = 0.009770655 - 0.001235232 = 0.008535423$$

y el intervalo de 95% de confianza para δ, el cual está dado por

$$\hat{\delta} - 1.96\text{ASE}(\hat{\delta}) \leq \delta \leq \hat{\delta} + 1.96\text{ASE}(\hat{\delta})$$

donde

$$\text{ASE}(\hat{\delta}) = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}};$$

aquí, $n_1 = 163858$ es la frecuencia total del grupo que no usó cinturón en sus accidentes, mientras que $n_2 = 412878$ es la frecuencia total del grupo que si lo utilizó. En el caso presente

$$\text{ASE}(\hat{\delta}) = 0.000249067$$

de manera que

$$0.008535423 - 1.96(0.000249067) \leq \delta \leq 0.008535423 + 1.96(0.000249067)$$

y finalmente,

$$0.008047253 \leq \delta \leq 0.009023594$$

enunciado que establece que, con un 95% de confianza, la diferencia de probabilidades se ubica entre 8 y nueve *milésimos*. Desde esta perspectiva, parece ser que π_1 y π_2 difieren por una cantidad insignificante, y que cualquier diferencia entre estas probabilidades es despreciable. Sin embargo, esta visión, como ya se apreció al analizar la tasa de riesgo y la razón de oportunidades, no es correcta, pues al comparar las magnitudes relativas de π_1 y π_2 se pone de manifiesto que π es alrededor de *ocho* veces π_2. □

4. Independencia y Análisis de Residuales

En el siguiente ejemplo se utilizan los procedimiento s X^2 y G^2 para probar independencia y, además, se introduce la idea de *residual estandarizado* para valorar la discrepancia entre lo observado y esperado en una tabla de contingencia.

Ejemplo 4.1. La siguiente tabla, tomada de la Encuesta Social General de 1991, realizada en Estados Unidos, se refiere a la identificación partidaria de las personas de acuerdo a su raza.

Identificación partidaria

Raza	Demócrata	Independiente	Republicano
1: Blanco	341	105	405
2: Negro	103	15	11

Con estos datos,

(a) Pruebe la independencia entre la identificación partidaria y la raza.

(b) Utilice los residuales ajustados para describir la evidencia.

Solución. Primeramente, y para própositos de referencia futura, se completará la tabla original anexando los total es de fila y columna, así como el gran total, indicado en negritas.

Raza	Demócrata	Independiente	Republicano	Total de filas
1: Blanco	341	105	405	851
2: Negro	103	15	11	129
Total de Columnas	444	120	416	**980**

(a) Se trata de probar la independencia entre las variables 'raza' y 'identificación partidaria', y para ello se utilizará tanto la prueba ji-cuadrada, com la prueba de razón de verosimilitud, en ambos casos con nivel de significancia de 5%.

(i) Prueba X^2 (ji-cuadrada): Recuerde que el estadístico X^2 está dado por

$$X^2 = \sum_{i\,j} \frac{(n_{i\,j} - \hat{\mu}_{i\,j})^2}{\hat{\mu}_{i\,j}}$$

donde $n_{i\,j}$ es la frecuencia en la celda i, j de la tabla original, mientras que $\hat{\mu}_{i\,j}$ es la frecuencia esperada bajo el supuesto de independencia, y está dada por

$$\hat{\mu}_{i\,j} = \frac{n_{i\,+} n_{+\,j}}{n_{+\,+}}$$

La siguente, es la tabla de frecuencias esperadas bajo el supuesto de independencia.

Raza	Demócrata	Independiente	Republicano	Total de filas
1: Blanco	385.555	104.2046	361.2408	851
2: Negro	58.4449	15.7959	54.7592	129
Total de Columnas	444	120	416	**980**

(4.1)

Tabla de Frecuencias Esperadas

A partir de esta nueva tabla y de la original, se determina la tabla de diferencias entre lo observado y lo esperado bajo el supuesto de independencia:

Raza	Demócrata	Independiente	Republicano	Total de filas
1: Blanco	−44.5551	0.7959	43.7591	0
2: Negro	44.5551	−0.7959	−43.7591	0
Total de Columnas	0	0	0	0

(4.2)

Tabla de Diferencias: Observado menos Esperado

Ahora los términos de X^2 se obtienen dividiendo el cuadrado de cada elemento de esta matriz, entre la componente correspondiente de la tabla precedente:

Raza	Demócrata	Independiente	Republicano
1: Blanco	5.148828552	0.006079282	5.300802316
2: Negro	33.96630309	0.040104414	34.96885869

Tabla de Términos de X^2

Por ejemplo,

$$\frac{(-44.55510204)^2}{385.555102} = 5.148828552$$

Sumando cada componente de esta última tabla se obtiene

$$X^2 = 79.43097634$$

Este estadístico tiene 2 grados de libertad, y el valor de probabilidad correspondiente a la observación 79.43 es, prácticamente, nulo—alrededor de 5×10^{-18}. Por lo tanto, la evidencia contra el supuesto de independencia entre las variables raza y afiliación partidista es concluyente.

(ii) Prueba G^2 (Razón de verosimilitud): El estadístico G^2 está dado por

$$G^2 = \sum_{i\,j} n_{i\,j} \log\left(\frac{n_{i\,j}}{\hat{\mu}_{i\,j}}\right)$$

donde $n_{i\,j}$ es la frecuencia en la celda i, j de la tabla original, mientras que $\hat{\mu}_{i\,j}$ es la frecuencia esperada bajo el supuesto de independencia. Por lo tanto, para calcular G^2 es natural proceder de acuerdo a los siguientes pasos: primero, se divide cada componente de la tabla original por el correspondiente miembro en la Tabla 4.1, se toma logaritmo natural al resultado y se multiplica este último valor por la frecuencia original. Se repite esto para cada entrada de la tabla original, obteniéndose la siguiente matriz:

Raza	Demócrata	Independiente	Republicano
1: Blanco	−83.75071988	1.597900598	92.61722245
2: Negro	116.7287862	−1.551041231	−35.31109604

Tabla de Términos de G^2

La suma de las componente de esta tabla proporciona el valor del estadístico G^2:

$$G^2 = 90.33105207$$

y el valor de probabilidad correspondiente a este número—bajo la distribución ji-cuadrada con dos grados de libertad—es prácticamente cero. Por lo tanto,, ante esta evidencia contundente, se rechaza el supuesto de independencia de las variables raza y afiliación partidista.

(ii) Como el supuesto de independencia ha sido rechazado, ahora se analizarán los residuales. Cada componente (i, j) tiene un residual (estandarizado), el cual se define como

$$r_{i\,j} = \frac{n_{i\,j} - \hat{\mu}_{i\,j}}{\sqrt{\hat{\mu}_{i\,j}(1 - \hat{\pi}_{i+})(1 - \hat{\pi}_{+j})}}$$

y tiene una distribución normal estándar bajo el supuesto de independencia, de manera que un valor absoluto de 2 o mas para el residual, indica que la hipótesis de independencia no es correcta.

Raza	Demócrata	Independiente	Republicano
1: Blanco	−8.456741218	0.229407788	8.364933114
2: Negro	8.456741218	−0.229407788	−8.364933114

Tabla de Residuales Estandarizados

En esta tabla hay cuatro residuales cuyo valor absoluto es mayor que ocho, y por lo tanto indican que la hipótesis de independencia entre las dos características estudiadas no es consistente con los datos observados. El valor −8.45 indica que

la frecuencia de personas blancas de afiliación demócrata es mucho menor a la esperada bajo el supuesto de independencia. Por lo tanto, en realidad las personas blancas se afilian menos al partido demócrata y más al republicano (como lo indica el residual de 8.5) en comparación a lo que se esperaría bajo independencia. Por el contrario, los residuales del grupo de raza negra indican que estas personas se afilian más al partido demócrata y menos al republicano que lo que se esperaría bajo el supuesto de independencia.

Como revela este análisis, una vez que se ha rechazado el supuesto de independencia, el estudio de los residuales es un instrumento relevante para dilucidar de donde proviene la asociación entre las variables involucradas en el estudio.

Nota: Es interesante obervar que en la tabla anterior los residuales en las columnas suman cero, mas no así, los de las filas. En general, cuando una variable tiene dos niveles—como la variable 'raza' en este ejemplo—la suma de los residuales estandarizados al sumarse sobre los niveles de dicha variable producirán el valor cero; esto sólo ocurre para una variable con dos niveles, como se observa al sumar los residuales estandarizados sobre las filas en la tabla precedente. □

Literatura Citada

[1]. Agresti A. 1996. Introduction to Categorical Data Analysis, Wiley, New York.

[2]. Agresti, A. 2004. Analysis of Ordinal Categorical Data. New York: Wiley.

[3]. Bishop, Y. M. M., S. E. Fienberg, y P. W. Holland. 1975. Discrete Multivariate Analysis. Cambridge, MA: MIT Press.

[4]. Brown, L. D., T. T. Cai, y A. Das Gupta. 2001. Interval estimation for a binomial proportion. Statist. Sci. 16: 101–133.

[5]. Dobson, A. J. 2001. An Introduction to Generalized Linear Models, 2nd ed. London: Chapman & Hall.

[6]. Ferguson, T. S. 1967. Mathematical Statistics: A Decision Theoretic Approach. New York: Academic Press.

[7]. Fleiss, J. L. 1981. Statistical Methods for Rates and Proportions, 2nd ed. New York: Wiley.

[8]. Goodman, L. A., y W. H. Kruskal. 1979. Measures of Association for Cross Classifications. New York: Springer-Verlag

[9]. Greenwood, P. E., y M. S. Nikulin. 1996. A Guide to Chi-Squared Testing. New York: Wiley.

[10]. Kendall, M., y A. Stuart. 1979. The Advanced Theory of Statistics, Vol. 2; Inference and Relationship, 4th ed. New York: Macmillan.

[11]. Khuri, A. I. 2002. Advanced Calculus with Applications in Statistics. New York: Wiley.

[12]. Pearson, K. 1922. On the χ^2 test of goodness of fit. Biometrika 14: 186–191.

[13]. Rao, C. R. 1973. Linear Statistical Inference and Its Applications, 2nd ed. New York: Wiley

[14]. Read, T. R. C., y N. A. C. Cressie. 1988. Goodness-of-Fit Statistics for Discrete Multivariate Data. New York: Springer-Verlag.

[15]. Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54: 426–482.