UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

"INTRODUCCIÓN A LOS MODELOS LOG - LINEALES"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A :

MARÍA EUGENIA SERRANO BARAJAS

DIRECTORA DE TESIS:

M. EN A.P. MARÍA DEL PILAR ALONSO REYES

FACULTAD DE CIENCIAS
UNAM

2005

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:
Introducción a los Modelos Log - Lineales

realizado por María Eugenia Serrano Barajas

con número de cuenta 9130131-0 , quien cubrió los créditos de la carrera de: Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	M. en A.P. María del Pilar Alonso Reyes
Propietario	M. en C. José Antonio Flores Díaz
Propietario	Act. María Aurora Valdés Michell
Suplente	Dr. Luis Antonio Rincón Solis
Suplente	Act. Jaime Vázquez Alamilla

Consejo Departamental de Matemáticas

Act. Jaime Vázquez Alamilla

AGRADECIMIENTOS

- A mi Directora de Tesis:

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

- A mis Sinodales:

M. en C. José Antonio Flores Díaz
Act. María Aurora Valdés Michell
Dr. Luis Antonio Rincón Solis
Act. Jaime Vázquez Alamilla

Índice

Introducción 1

Capítulo 1
Características probabilísticas de las variables aleatorias **3**

1.1 Variable aleatoria discreta 3
1.2 Función de distribución de una variable aleatoria discreta 3
1.3 Probabilidad condicional e independencia 5
1.4 Características de la variable aleatoria discreta 5
Esperanza 5
Varianza 7
Desviación estándar 7
1.5 Variable aleatoria discreta conjunta 8
1.6 Función de distribución de probabilidad de una variable discreta conjunta 8
1.7 Probabilidad condicional e Independencia 9
Teorema de Bayes 10
1.8 Funciones marginales de densidad de probabilidad 12
1.9 Características de las variables aleatorias discretas conjuntas 13
Esperanza 13
Covarianza 14
Coeficiente de correlación 14
Matriz de varianzas y covarianzas 15

Capítulo 2
Algunas densidades importantes para los Modelos Log-Lineales **16**

2.1 La distribución binomial 16
2.2 La distribución multinomial 20
2.3 La distribución de producto de multinomiales 23
2.4 La distribución Poisson 24

Capítulo 3
Tablas de contingencias **29**

3.1 Introducción 29
3.2 Distribuciones de frecuencias 32
3.3 Tablas de contingencias 33
3.4 Tipos de tablas de contingencias 35
3.5 Distribuciones de más de dos dimensiones 35
3.6 Probabilidad condicional e independencia 38

Capítulo 4
Modelos Log-Lineales **42**

4.1 Tablas de 2x2 42
4.2 Prueba de independencia para una tabla de 2x2 44
4.3 Tablas de CxR 50
4.4 Teoría de máxima verosimilitud para tablas de dos dimensiones 54
4.5 Modelos Log-Lineales para tablas de dos dimensiones 61

Capítulo 5
Aplicación del Modelo Log-Lineal a los siniestros de una cartera de seguros de vida individual **74**

5.1 Introducción 74
5.2 Estadísticas descriptivas 79
5.3 Resultados del Modelo Log-Lineal de 2x2 83
5.4 Resultados del Modelo Log-Lineal de 3x3 88

Conclusiones **97**
Bibliografía **100**

Introducción

Los modelos log-lineales representan elementos útiles para el análisis de datos que pueden clasificarse en distintas categorías, como pueden ser el caso de encuestas, considerando las diferentes variables que constituyen un modelo, pueden ayudar a determinar si existe relación entre las variables involucradas.

Para el análisis de la información de un modelo clasificado en categorías, basados en la teoría de los modelos log-lineales, se considera el logaritmo de cada de una de las celdas de la tabla de frecuencias generada de la información disponible y, considerando las características de los diferentes tipos de modelos log-lineales que existen, se decide qué modelo se ajusta de la mejor manera a dicha información y sus características.

En el desarrollo del presente trabajo, se pretende tener una herramienta que facilite el análisis de información que se pueda clasificar en categorías y forme una tabla de contingencias, posteriormente esta teoría se aplica a una información real, detallando los modelos y el proceso de ajuste a la información. Para llegar a este objetivo se comienza en el capítulo 1 revisando algunos conceptos de probabilidad y estadística para variables aleatorias discretas, que serán necesarios para desarrollar los siguientes temas, como son media, varianza, coeficiente de correlación, pruebas de independencia, etc.

En el capítulo 2, se estudian algunas distribuciones comunes en el análisis de los modelos log-lineales, como son la binomial, multinomial y Poisson, así como sus principales características. En el capítulo 3, se presenta una introducción al estudio de las tablas de contingencias sus características y propiedades, además de los tipos de frecuencias y los tipos de tablas de contingencias.

En el capítulo 4 se integra toda la teoría referente a los modelos log-lineales, comenzando desde el análisis del modelo de dimensión 2×2, y generalizando a dimensión $C\times R$, revisando la independencia en las variables que forman las tablas de contingencias con pruebas de hipótesis y obteniendo los estimadores de máxima verosimilitud de los modelos.

En el capítulo 5 se ajusta un modelo log-lineal a una cartera de seguros de vida individual, considerando los siniestros ocurridos en un periodo determinado en todo el país. Se pretende analizar si variables como sexo y edad influyen sobre la variable causa de muerte. Se revisa un modelo con dos variables y posteriormente se agrega una variable más. Al final se presentan las conclusiones.

Capítulo I

Características probabilísticas de las variables aleatorias

En este capítulo se revisarán conceptos básicos de probabilidad y estadística que serán necesarios para el desarrollo posterior. La mayoría de estos conceptos fueron tomados de libros como Introducción a la Teoría de la Estadística y Estadística Matemática con Aplicaciones, ambos citados en la bibliografía.

1.1 Variable aleatoria discreta

Se dice que X es una **variable aleatoria discreta** si toma sólo un número finito o infinito numerable de valores del eje x[1]. Esta variable corresponde a experimentos en los que se cuenta el número de veces que ha ocurrido un suceso.

1.2 Función de distribución de una variable aleatoria discreta

La distribución de probabilidad de una variable aleatoria discreta X se puede representar con una tabla, fórmula o gráfica que indique las probabilidades $p(x)$ correspondientes a cada uno de los valores de x.

Un modelo de distribución de probabilidad es la representación idealizada de un experimento aleatorio y se construye indicando los posibles valores de la variable aleatoria asociada al experimento y sus probabilidades respectivas. La forma más

[1] P. 61 Mood and Graybill

general de caracterizar estos modelos es mediante la función de distribución, $F(x)$, definida en cada punto x_0 como la probabilidad de que una variable aleatoria X tome un valor menor o igual que x_0, lo cual se escribe:

$$F(x_0) = P(X \leq x_0).$$

La función de distribución, se define para todo el eje real, es siempre no decreciente, y por definición:

$$F(-\infty) = 0$$

$$F(+\infty) = 1$$

Si se supone que la variable X toma todos los valores posibles $(x_1,...,x_n)$, siendo

$$x_1 \leq x_2 \leq ... \leq x_n,$$

$$\sum P(x_i) = 1,$$

entonces la función de distribución acumulada será definida por:

$$F(x_1) = P(X \leq x_1) = P(x_1),$$

$$F(x_2) = P(X \leq x_2) = P(x_1) + P(x_2),$$

$$\text{N} \qquad \text{N} \qquad \text{N} \qquad \text{N}$$

$$F(x_n) = P(X \leq x_n) = \sum_{i=1}^{n} P(x_i) = 1.$$

La función de distribución $F(x)$ tendrá saltos en los puntos $(x_1,...,x_n)$ iguales a la probabilidad de dicho punto, siendo constante en los intervalos entre los puntos del salto.

1.3 Probabilidad condicional e independencia

La probabilidad condicional y la probabilidad marginal juegan un papel importante en el análisis de los modelos log-lineales. Si *A* y *B* son dos sucesos de un espacio muestral *S* tal que P(*B*)>0. La probabilidad condicional del suceso *A* dado el suceso *B* se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

esta es la proporción de la probabilidad de *A* dado el caso de que *B* ocurrió[2].

Si se conoce que el evento *B* ocurre y no cambia la información acerca de *A*, entonces se dice que *A* es independiente de *B*; específicamente, *A* es independiente de *B* sí:

$$P(A|B) = P(A).$$

Esta definición está condicionada a que *P(B)>0*. Una definición simple y equivalente es:

$$P(A \cap B) = P(A)P(B).$$

1.4 Características de la variable aleatoria discreta

La esperanza de una variable aleatoria es un número que caracteriza la media o valor esperado de la distribución. Para una variable aleatoria Y, con una distribución discreta, el valor esperado está definido como:

[2] P. 41 Mood & Graybill

$$E(Y) = \sum_{\forall r} rP(Y = r).$$

Tómese nota que distribuciones con el mismo valor esperado, pueden ser muy diferentes entre si debido a que la esperanza de una variable aleatoria sólo indica el centro de la distribución, pero no la variación o dispersión que tiene la variable alrededor de la media. Más adelante se definirá una medida de dispersión.

Teorema: Sea Y una variable aleatoria discreta con una función de probabilidad $p(y)$ y sea $g(y)$ una función de valores reales de Y. Entonces el valor esperado de $g(y)$ está dado por:

$$E[g(y)] = \sum_{y} g(y)p(y).$$

Teorema: Sea c una constante, entonces $E(c) = c$.

El siguiente teorema establece que el valor esperado del producto de una constante *c* por una función de una variable aleatoria es igual al producto de la constante por el valor esperado de la función.

Teorema: Sea $g(Y)$ una función de una variable aleatoria Y y sea c una constante. Entonces:

$$E[cg(Y)] = cE[g(Y)].$$

El teorema siguiente, establece que la media o el valor esperado de la suma de funciones de una variable aleatoria Y es igual a la suma de sus respectivos valores esperados.

Teorema: Sean $g_1(Y), g_2(Y), ..., g_k(Y)$ funciones de la variable aleatoria Y. Entonces:

$$E[g_1(Y) + g_2(Y) + ... + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + ... + E[g_k(Y)].$$

La **varianza** de una distribución es en realidad una medida de dispersión definida como el valor esperado del cuadrado de las desviaciones de la variable aleatoria con respecto a su media o valor esperado. Sea *Y* una variable aleatoria, entonces la varianza de *Y* se define como :

$$Var(Y) = \sum_{\forall r} (r - \mu)^2 P(Y = r).$$

Teorema: $Var(Y) = \sigma^2 = E\left[(Y - \mu)^2\right] = E(Y^2) - \mu^2$

Este teorema reduce considerablemente la tarea de encontrar la varianza de una variable aleatoria discreta.

El problema con la varianza es que es una medida con una mala escala. Por ejemplo, si *Y* es una variable cuya una unidad está dada en metros, $Var(Y)$ por definición envuelve el término $(y - \mu)^2$, de aquí que sea una medida en metros cuadrados. Para tener estos términos en escala comparable de los datos originales, se considera a la **desviación estándar** de Y que se define como:

$$DE(Y) = \sigma = \sqrt{Var(Y)}\,.$$

Tanto la varianza y desviación estándar son utilizadas como medidas de dispersión relativa y se utiliza en la comparación de dos conjuntos de mediciones. Ambas tienen particular importancia por ser convenientes matemáticamente y al ser utilizadas comúnmente en la distribución normal (gaussiana), la que está totalmente caracterizada por su valor esperado (media) y varianza. Por otra parte, existen otras medidas igualmente buenas de dispersión que pueden dar resultados inconsistentes con los datos.

1.5 Variable aleatoria discreta conjunta

En muchas ocasiones, los fenómenos que se tratan de investigar, pueden incluir valores de varias variables relacionadas entre sí, por lo que es interesante su estudio en conjunto. Se supondrá que sobre cada elemento o individuo se han observado varias variables, en lugar de una.

Variable aleatoria discreta conjunta: La variable aleatoria de dimensión *k*, $(X_1, X_2, ..., X_k)$, es definida discreta si sólo puede tomar valores en un conjunto finito o infinito numerable de puntos $(x_1, x_2, ..., x_k)$ en el espacio real de dimensión k^3.

1.6 Función de distribución de probabilidad de una variable aleatoria discreta conjunta

Sea $(X_1, X_2, ..., X_k)$ una variable aleatoria discreta conjunta de dimensión *k*, **la función de probabilidad conjunta discreta** de $(x_1, x_2, ..., x_k)$ es:

$$f_{x_1,\ldots,x_k}(X_{1,\ldots,}X_k)=\begin{cases} p(X_1=x_1,X_2=x_2,\ldots,X_k=x_k) & \text{Para } (x_1,x_2,\ldots,x_k) \\ 0 & \text{Para cualquier otro caso} \end{cases}$$

1.7 Probabilidad condicional e independencia

Distribuciones condicionales

Sea I un conjunto no vacío de subconjuntos de los enteros $\{1,2,\ldots,n\}$. Si $f(x_1,x_2,\ldots,x_n)$ es una función de densidad de probabilidad discreta, se tiene:

$$p\left[X_{j_1}=x_{j_1},X_{j_2}=x_{j_2},\ldots,X_{j_{n-k}}=x_{j_{n-k}}\middle|X_{i_1}=x_{i_1},X_{i_2}=x_{i_2},\ldots,X_{i_k}=x_{i_k}\right]$$

$$=\frac{p(X_1=x_1,X_2=x_2,\ldots,X_n=x_n)}{p(X_{i_1}=x_{i_1},X_{i_2}=x_{i_2},\ldots,X_{i_k}=x_{i_k})},$$

donde $I=\{i_1,i_2,\ldots,i_k\}$ y $I^c=\{j_1,j_2,\ldots,j_{n-k}\}$, entonces:

$$f\left(X_{j_1},X_{j_2},\ldots,X_{j_{n-k}}\middle|x_{i_1},x_{i_2},\ldots,x_{i_k}\right)=\frac{f(x_1,x_2,\ldots,x_n)}{f\left(x_{i_1},x_{i_2},\ldots,x_{i_k}\right)}$$

es la función de probabilidad marginal discreta.

Independencia

Como se analizó al principio del texto dos variables aleatorias son independientes si el conocimiento de una de ellas no aporta información

[3] Pag. 69 Mood & Graybill

respecto a los valores de la otra. Esta definición se extiende a cualquier conjunto de variables aleatorias .

Si se tienen *n* variables aleatorias, éstas serán mutuamente independientes entre si:

$$p(a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2, \ldots, a_n < x_n \leq b_n) = \prod_{i=1}^{n} p(a_i < x_i \leq b_i).$$

La función de distribución acumulativa será:

$$\prod_{i=1}^{n} F_{x_i}(b_i)$$

cuando $a_i \to -\infty$ para todo *i*.

Teorema de Bayes

Considérese un experimento que se realiza en dos etapas: en la primera, los posibles sucesos $a_1, \ldots, a_n$ son mutuamente excluyentes, con probabilidades conocidas $P(a_i)$, tales que:

$$\sum_{i=1}^{n} P(a_i) = 1$$

En la segunda etapa, los resultados posibles b_j, dependen de los de la primera, y se conocen las probabilidades condicionadas $P(b_j|a_i)$ de obtener cada posible resultado b_j cuando ocurre en la primera etapa el a_i.

Se efectúa el experimento, pero el resultado de la primera fase, a_i, no se conoce, aunque si el de la segunda, que resulta ser b_j. El Teorema de Bayes

permite calcular las probabilidades $p(a_i|b_j)$ de los sucesos no observados en la primera etapa, dado el resultado de la segunda.

Partiendo de la definición de probabilidad condicionada:

$$p(a_i|b_j) = \frac{P(a_i \cap b_j)}{P(b_j)} = \frac{P(b_j|a_i)P(a_i)}{P(b_j)},$$

por otro lado:

$$P(b_j) = P(b_j a_1 \cup b_j a_2 \cup ... \cup b_j a_n)$$

ya que b_j debe ocurrir con alguno de los n posibles eventos a_i. Como los sucesos $b_j a_1, ..., b_j a_n$ son mutuamente excluyentes, al serlo los a_i, se obtiene que:

$$P(b_j) = \sum_{i=1}^{n} P(b_j \cap a_i) = \sum_{i=1}^{n} P(b_j|a_i)P(a_i),$$

sustituyendo en la expresión de $p(a_i|b_j)$ se logra:

$$p(a_i|b_j) = \frac{P(b_j|a_i)P(a_i)}{\sum_{i=1}^{n} P(b_j|a_i)P(a_i)}$$

que se conoce como el Teorema de Bayes.

La idea de la probabilidad condicional admite una generalización inmediata a situaciones donde interviene más de un criterio de clasificación; por ejemplo en el caso de tres etapas se ve claramente que:

$$P(a_i \cap b_j | c_k) = \frac{P(a_i \cap b_j \cap c_k)}{P(c_k)}$$

$$P(a_i | b_j \cap c_k) = \frac{P(a_i \cap b_j \cap c_k)}{P(b_j \cap c_k)},$$

y también que:

$$\begin{aligned} P(a_i \cap b_j \cap c_k) &= P(a_i \cap b_j | c_k) P(c_k), \\ &= P(a_i | b_j \cap c_k) P(b_j \cap c_k), \\ &= P(a_i | b_j \cap c_k) P(b_j | c_k) P(c_k). \end{aligned}$$

Pueden obtenerse otras relaciones análogas permutando las letras *a*,*b* y *c*. Así:

$$P(b_j | a_i \cap c_k) = \frac{P(a_i \cap b_j \cap c_k)}{P(a_i \cap c_k)}$$

y

$$P(a_i \cap b_j \cap c_k) = P(b_j | a_i \cap c_k) P(a_i | c_k) P(c_k),$$

o bien

$$P(a_i \cap b_j \cap c_k) = P(b_j | a_i \cap c_k) P(c_k | a_i) P(a_i).$$

1.8 Funciones marginales de densidad de probabilidad

Así se denomina a la distribución de la variable considerada aisladamente, con independencia de las demás. Las funciones de densidad marginales para cada X_i con $i = 1,2,...,k$ serán:

$$f(x_1) = \sum_{x_k} \sum_{x_{k-1}} ... \sum_{x_2} f(x_1, x_2, ..., x_k),$$

N

$$f(x_k)=\sum_{X_{k-1}}\sum_{X_{k-2}}\cdots\sum_{X_1} f(x_1,x_2,\ldots,x_k).$$

1.9 Características de las variables aleatorias discretas conjuntas

Esperanza en una variable de dimensión *n*

Sea $h(X_1,X_2,\ldots,X_n)$ alguna función del vector $(X_1,X_2,\ldots,X_n)$ entonces:

$$E(h(X_1,X_2,\ldots,X_n))=\sum_{X_n}\cdots\sum_{X_1} h(x_1,x_2,\ldots,x_n)f(x_1,x_2,\ldots,x_n)$$

para el caso discreto.

La esperanza de un vector aleatorio $\underline{X}=(X_1,X_2,\ldots,X_n)$, es aquel cuyos componentes son las esperanzas de cada una de variables X_i y cuya notación es:

$$\underline{\mu}=E[\underline{X}]$$

donde la **esperanza** de un vector **o** una matriz debe entenderse como el resultado de aplicar este operador (tomar medias) a cada uno de sus componentes.

Esperanza de sumas y productos

Dadas n variables aleatorias definidas conjuntamente con una función de densidad $f(X_1,\ldots,X_n)$, se verifica que:

$$E(X_1 + X_2 + ... + X_n) = E[X_1] + E[X_2] + ... + E[X_n]$$

Para variables independientes $f(X_1,...,X_n) = f(X_1)f(X_2)...f(X_n)$:

$$E(X_1 \cdot X_2 \cdot ... \cdot X_n) = E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i]$$

La esperanza de un producto es el producto de las esperanzas.

Teorema: La **covarianza** es una medida de relación o dependencia lineal entre dos variables aleatorias. Para definirla, se suponen dos variables aleatorias Y_1 y Y_2, donde $E(Y_1) = \mu_1$ y $E(Y_2) = \mu_2$. La covarianza entre Y_1 y Y_2 se da por:

$$Cov(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = E(Y_1 Y_2) - \mu_1 \mu_2$$

Por otro lado puede observarse que la varianza puede interpretarse como:

$$Var(Y_1) = Cov(Y_1, Y_1)$$

En un intento por saber cómo obtener un valor numérico que no dependa de la escala de medición de la característica se considera el **coeficiente de correlación,** dado por la expresión:

$$Corr(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}}$$

Una perfecta relación lineal creciente está indicada por el valor de 1, una decreciente por -1 y la ausencia de relación lineal está dada por cero.

Matriz de varianzas y covarianzas

Se define la matriz de varianzas y covarianzas de un vector aleatorio $\underline{X}$ a la matriz cuadrada de orden n:

$$\mathrm{M}_x = E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})']$$

Llamando a $\underline{X} = (X_1,...,X_n)$, $\underline{\mu} = (\mu_1,...,\mu_n)$, se obtiene que la matriz M_x contiene en la diagonal las varianzas de los componentes y fuera de ella las covarianzas entre dos variables cualesquiera. La matriz M_x será siempre simétrica y semidefinida positiva, es decir, todos los menores principales serán positivos y dado un vector de números cualesquiera ω se verificará que:

$$\underline{\omega}' \mathrm{M}_x \underline{\omega} \geq 0$$

Esta propiedad se comprueba definiendo una variable unidimensional por:

$$v = (\underline{X} - \underline{\mu})' \cdot \underline{\omega},$$

y como la varianza de v debe ser no negativa:

$$\mathrm{var}(v) = E\left[v^2\right] = \underline{\omega}' E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})']\underline{\omega} \geq 0$$

Capítulo 2

Algunas densidades importantes para los Modelos Log-Lineales

Para el desarrollo de los modelos log-lineales, existen algunas funciones de distribución que se utilizan comúnmente. En este capítulo se repasarán algunas de ellas y sus principales propiedades.

2.1 La distribución binomial

Si se tienen *n* eventos independientes y se está contando qué tan seguido un evento ocurre, el de fracaso o no fracaso, la distribución apropiada para modelar tal situación es la binomial. Típicamente, el resultado de interés es referido como un éxito, cuya probabilidad se denota por *p*, para cada uno de los *n* ensayos, entonces *X*, el número de éxitos, tiene una distribución binomial con parámetros *n* y *p*. La distribución en cuestión se puede denotar como:

$$X \sim Bin(n, p).$$

Que analíticamente se expresa por:

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r},$$

r =1,...,n y donde:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Con base en lo anterior se puede encontrar la media (valor esperado) y la varianza. Por definición, la media está dada por:

$$E(X) = \sum_{r=0}^{n} r \binom{n}{r} p^r (1-p)^{n-r} = np$$

La varianza de *X* está definida como:

$$Var(X) = \sum_{r=0}^{n} (r-np)^2 \binom{n}{r} p^r (1-p)^{n-r} = np(1-p)$$

En algunas ocasiones se necesita trabajar con el número de éxitos y el de fracasos al mismo tiempo. Si los primeros se refieren como X_1 y los segundos como X_2, entonces:

$$X_2 = n - X_1,$$

$$X_1 \sim \textit{Bin(n, p)} \text{ y } X_2 \sim \textit{Bin(n, 1-p)}.$$

Este último resultado se apoya en el hecho de que con ensayos idénticos independientes, el número de resultados que se puede llamar fracasos deben tener también una distribución binomial. Si la probabilidad de éxito es *p*, la probabilidad de fracaso es *1-p*. Por supuesto:

$$E(X_2) = n(1-p)$$

y

$$Var(X_2) = n(1-p)p$$

Es importante notar que:

$$Var(X_1) = Var(X_2)$$

Es indiferente del valor de *p*. Finalmente:

$$Cov(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2]$$

donde $E[X_1] = np$, $E[X_2] = n(1-p)$ y $E[X_1 X_2] = np(1-p)(n-1)$, entonces se obtiene que:

$$Cov(X_1, X_2) = -np(1-p)$$

El coeficiente de correlación está definido como:

$$Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2}.$$

Donde σ_1 y σ_2 son las desviaciones estándar, que por definición son $\sqrt{Var(X_1)}$ y $\sqrt{Var(X_2)}$. Sustituyendo:

$$Corr(X_1, X_2) = \frac{-np(1-p)}{\sqrt{np(1-p)}\sqrt{n(1-p)p}}.$$

Recordando que $Var(X_1) = Var(X_2)$

$$Corr(X_1, X_2) = \frac{-np(1-p)}{(\sqrt{np(1-p)})^2},$$

$$= \frac{-np(1-p)}{(\sqrt{np(1-p)})^2},$$

$$= \frac{-np(1-p)}{np(1-p)},$$

=-1.

Lo anterior se puede interpretar como que existe una perfecta relación lineal entre X_1 y X_2. Si X_1 aumenta una unidad, X_2 disminuye una unidad.

Cuando se observan ambos tipos de eventos, éxitos y fracasos, entonces:

$(X_1, X_2) \sim$ Bin(n, p,(1-p)).

2.2 La distribución multinomial

La distribución multinomial es una generalización de la binomial para más de dos categorías. Suponiendo que se tienen *n* ensayos idénticos e independientes. En cada ensayo se observa que hay *q* eventos que pueden ocurrir. En cada uno de los ensayos se asume que, uno de los *q* eventos debe ocurrir. Sea X_i, *i*=1,2,...,*q*, el número de veces que el *i*-ésimo evento ocurre. Sea p_i la probabilidad de que el *i*-ésimo evento ocurra en algún ensayo. Hay que hacer notar que las p_i deben satisfacer $p_1 + p_2 + p_3 ... + p_q = 1$ En este caso se dice que $(X_1, X_2, X_3, ..., X_q)$ tienen una distribución multinomial con parámetros $(n, p_1, p_2, p_3, ..., p_q)$. Tal como a continuación se escribe: $(X_1, X_2, X_3, ..., X_q) \sim Mult(n, p_1, p_2, p_3, ..., p_q)$.

La distribución es:

$$\Pr(X_1 = r_1, ..., X_q = r_q) = \frac{n!}{r_1! .. r_q!} p_1^{r_1} ... p_q^{r_q} = \frac{n!}{\prod_{i=1}^{q} r_i!} \prod_{i=1}^{q} p_i^{r_i} .$$

Para $r_i \geq 0$ y $r_1 + r_2 + ... + r_q = n$. Si *q*=2, ésta es la distribución binomial. En general, cada componente individual es:

$$X_i \sim \text{Bin}(n, p_i).$$

Teorema: Si $X_1, X_2, ..., X_q$ tienen una distribución multinomial con parámetros n y $p_1, p_2, ..., p_q$, entonces

$$E(X_i) = n(p_i),$$

$$Var(X_i) = n(1 - p_i)p_i,$$

$$Cov(X_i, X_j) = -np_ip_j \text{ para } i \neq j.$$

Demostración:

Se puede usar la distribución marginal de X_i para obtener la media y la varianza. Puede interpretarse a X_i como el número de pruebas que caen en una casilla *i*.

Hay que considerar todas las casillas combinadas en una casilla grande, menos la casilla *i*. Así cada prueba tendrá su resultado en la casilla *i* o en una casilla diferente a la *i* con probabilidades p_i y $1 - p_i$, respectivamente. Por lo tanto X_i posee una distribución de probabilidad marginal binomial. Por consiguiente:

$$E(X_i) = n(p_i) \text{ y } Var(X_i) = n(1 - p_i)p_i.$$

Para la demostración de $Cov(X_i, X_j) = -np_ip_j$ para $i \neq j$, se considerará el experimento multinomial como una sucesión de *n* pruebas independientes, se define:

$$U_t = \begin{cases} 1, & \text{Si la prueba } t \text{ cae en la casilla } i. \\ 0, & \text{En caso contrario.} \end{cases}$$

$$W_j = \begin{cases} 1, & \text{Si la prueba s cae en la casilla } j. \\ 0, & \text{En caso contrario.} \end{cases}$$

Entonces $X_i = \sum_{t=1}^{n} U_t$ y $X_j = \sum_{s=1}^{n} W_s$.

Hay que observar que X_i es una suma de ceros y unos porque U_t=1,0 cuando la t-ésima prueba cae en la casilla *i* o no. Hay un 1 en la suma por cada vez que se observa un elemento de la clase *s* y un 0 para cada vez que se tome cualquier otra clase. Así, X_i es el número de veces que se observa la clase *i*. Puede hacerse una interpretación similar de X_j. Para evaluar la $Cov(X_i, X_j)$, se necesitan los resultados siguientes:

$$E(U_i) = P_i,$$

$$E(W_j) = P_j ,$$

$Cov(U_s, W_t) = 0$ si $s \neq t$ ya que las pruebas son independientes,

$$Cov(U_s, W_t) = E(U_s W_t) - E(U_s)E(W_t) = 0 - p_s p_t .$$

Porque $U_s W_t$ siempre es igual a cero. Se utilizará también el teorema que dice lo siguiente:

Sean $X_1, X_2, ..., X_n$ y $Y_1, Y_2, ..., Y_m$ variables aleatorias con $E(X_i) = \mu_i$ y $E(Y_i) = \xi_i$. Defina $U_1 = \sum_{i=1}^{n} a_i x_i$ y $U_2 = \sum_{j=1}^{m} b_j y_j$, para constantes $a_1, a_2, ..., a_n$ y $b_1, b_2, ..., b_m$, entonces se cumple que $Cov(U_1, U_2) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)$. Por consiguiente:

$$Cov(X_s, X_t) = \sum_{s=1}^{n}\sum_{t=1}^{n} Cov(U_s, W_t),$$

$$= \sum_{s=1}^{n} Cov(U_s, W_s) + \sum_{s \neq t}\sum Cov(U_s, W_t),$$

$$= \sum_{s=1}^{n}(-p_t p_j) + 0 = -np_t p_j .$$

Se observa que la covarianza es negativa, lo que se esperaba debido a que un gran número de resultados en la casilla *i* implica un pequeño número de resultados en la casilla *j* y viceversa.

2.3 La distribución producto de multinomiales

Para *i*=1,...t, se toman multinomiales independientes donde la *i*-ésima tiene s_i posibles resultados, es decir, $(X_{i1}, X_{i2}, ..., X_{is}) \sim \text{Mult}(N_i, p_{i1}, p_{i2}, ..., p_{is})$, entonces se dice que las x_{ij} tienen una distribución producto de multinomiales. Por independencia, la probabilidad de cualquier conjunto de resultados, es decir $\Pr(X_{ij} = r_{ij})$ para todo *i,j* es el producto de las probabilidades mutinomiales para cada *i*. En otras palabras:

$$\Pr(X_{ij} = r_{ij}) = \prod_{i=1}^{t} \Pr(X_{ij} = r_{ij}) \text{ para todo } i,j,$$

y para $r_{ij} \geq 0$, $j = 1,...,s_i$ con $\sum_{j=1}^{s_i} r_{ij} = N_i$, se tiene:

$$\Pr(X_{ij} = r_{ij}) = \left(\frac{N_i!}{\prod_{j=1}^{s_i} r_{ij}!} \right) \prod_{j=1}^{s_i} (p_{ij})^{r_{ij}} .$$

Entonces:

$$\Pr(X_{ij} = r_{ij}) = \prod_{i=1}^{t} \left(\frac{N_i!}{\prod_{j=1}^{s_i} r_{ij}!} \right) \prod_{j=1}^{s_i} (p_{ij})^{r_{ij}} .$$

Dados $r_{ij} \geq 0$ para todo i,j y $r_{i1} + r_{i2} + ... + r_{is_i} = N_i$ para toda *i*. Medias, varianzas y covarianzas sin una multinomial particular es obtenida ignorando a otras multinomiales.

Las covarianzas entre puntajes en multinomiales diferentes es cero porque las observaciones son independientes.

2.4 La distribución Poisson

Las distribuciones binomial y multinomial son muy usadas y apropiadas cuando el número de ensayos no es muy grande (cualquiera que sea la media) y las probabilidades de los sucesos no son muy pequeñas. Para un fenómeno que tiene

una probabilidad muy pequeña de ocurrir en un ensayo particular, pero para el cual, un número extremadamente grande de ensayos es plausible, la distribución Poisson es apropiada. Por ejemplo, el número de suicidios en un año debería tener una distribución Poisson ya que, la probabilidad de que alguna persona se suicide es muy pequeña, pero en una población grande, un número sustancial de personas podrían hacerlo.

La distribución Poisson puede ser obtenida como el límite de una *Bin(n,p)*, cuando $n \to \infty$ y $p \to 0$. De cualquier modo, las convergencias deberían ocurrir en caso de que $np \to \lambda$. El valor de λ es el parámetro de la distribución Poisson. Si *X* es una variable aleatoria con distribución Poisson y parámetro λ:

$$X \sim \text{Poisson}(\lambda) \qquad (1)$$

La distribución de probabilidades se define como:

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}, \text{ para } r = 0,1,\ldots$$

Es fácil llegar a (1) observando la probabilidad binomial correspondiente para *X=r* es:

$$\binom{n}{r} p^r (1-p)^{n-r} = \left[\frac{(np)^r (1-p)^n}{r!}\right](1-p)^{-r}\frac{n!}{(n-r)!n^r} \qquad (2)$$

con $n \to \infty$ y $p \to 0$ y $np \to \lambda$,

$$(np)^r \to \lambda^r$$

$$(1-p)^n \to e^{-\lambda}$$

$$(1-p)^{-r} \to 1$$

$$\frac{n!}{(n-r)!n^r} \to 1$$

sustituyendo estos límites en el lado derecho de (2), da la forma de la probabilidad en (1).

Demostración de la convergencia de la función de probabilidad binomial hacia la Poisson

$$\lim_{n\to\infty}\binom{n}{r}p^r(1-p)^{n-r} = \lim_{n\to\infty}\frac{n(n-1)\ldots(n-r+1)}{r!}\left(\frac{\lambda}{n}\right)^r\left(1-\frac{\lambda}{n}\right)^{n-r},$$

$$= \lim_{n\to\infty}\frac{\lambda^r}{r!}\left(1-\frac{\lambda}{n}\right)^n\frac{n(n-1)\ldots(n-r+1)}{n^r}\left(1-\frac{\lambda}{n}\right)^{-r},$$

$$= \frac{\lambda^r}{r!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-r}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right).$$

Sabiendo que $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$

y que los demás términos a la derecha de la expresión anterior como límite tienen límite a 1, se obtiene que:

$$p(X) = \frac{\lambda^r}{r!}e^{-\lambda}.$$

Utilizando (1), puede calcularse el valor esperado y la varianza de *X*. No es difícil ver que:

$$E(X) = \lambda$$

y

$$Var(X) = \lambda .$$

Por otra parte, si $X_1, X_2, ..., X_q$ son variables aleatorias independientes con distribución Poisson(λ_i) cada una, entonces la suma de ellas se distribuye Poisson con parámetro igual a la suma de los parámetros de cada X_i, lo anterior se considera de la siguiente manera:

Si X_i ~Poisson(λ_i) para toda i=1,...,n entonces:

$$X_1 + X_2 + ... + X_q \sim \text{Poisson}(\lambda_i + \lambda_2 + ... + \lambda_q),$$

y las cantidades dadas en el total son:

$$(X_1 + X_2 + ... + X_q) | N \sim \text{Mult}(N, p_1, p_2, ..., p_q).$$

Donde $N = (X_1 + X_2 + ... + X_q)$ y $p_i = \dfrac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_q}$, i=1,2,...,q.

La distribución condicional es importante para el análisis de los modelos log-lineales. Si se tiene una tabla de puntajes que está formada por variables aleatorias independientes Poisson, se puede calcular siempre el gran total de la tabla. Observando la distribución condicional dando el total de salidas, da un

Capítulo 3

Tablas de contingencias

Una vez revisados los temas básicos para el seguimiento del tema, se comenzará a estudiar la teoría correspondiente a las tablas de contingencias y sus características. Este capítulo se apoyó en los libros *Log-Linear Models*, Estadística Modelos y Métodos.

3.1 Introducción

Es posible clasificar a los miembros de una población en términos genéricos denotando alguna buena definición para la clase de personas o cosas en formas muy variadas. Las personas, en primera instancia, pueden catalogarse como femenino o masculino, casada o soltera, en aquellas que optan por votar en las elecciones y las que se abstienen de hacerlo. Estos son algunos ejemplos de clasificaciones dicotómicas. Existen también categorizaciones comunes como cuando se divide a la gente en zurdo, ambidiestro, derecho; otra clasificación común es en las elecciones, al votar, a) priísta, b) panista, c) perredista, d) aquellos indecisos y e) otros. Las clasificaciones que particularmente interesan son las exhaustivas y mutuamente excluyentes. Una clasificación es exhaustiva cuando provee de categorías suficientes para acomodar a todos los miembros de una población. Las categorías son mutuamente excluyentes cuando son definidas de tal manera que cualquier miembro de la población puede ser correctamente clasificado en una y solamente una categoría. A primera vista, puede parecer que los requerimientos para que una clasificación sea exhaustiva son muy restrictivos. Por ejemplo, si se estuviera interesado en llevar a cabo una encuesta, no referente a las preferencias de votar de los electores en su

totalidad, pero sí en las de los estudiantes universitarios. La dificultad es resolver si la definición de la población es adecuada, la definición estadística de las palabras es mucho más amplia que la usual, pero sin embargo, es totalmente correcto definir a la población del ejemplo como " todos los estudiantes universitarios que tienen derecho a votar". Hay categorías que son muy ajustables y frecuentemente pueden ofrecerse modificadas o combinadas; en el ejemplo de la votación es poco probable que pueda perderse mucha información por fusionarse las categorías d) (aquellos indecisos) y e) (otros).

Cuando la población es clasificada dentro de varias categorías, se puede "contar" el número de individuos en cada una de ellas. Estos "puntajes" o frecuencias son el tipo de datos que son el interés principal y se debe tratar con datos cualitativos en lugar de datos cuantitativos, obtenidos de medidas de variables continuas tales como peso, estatura, etcétera.

En general, la información de la población entera no está disponible y entonces se debe trabajar con una muestra de la población. Una función de la Estadística es demostrar como inferencias válidas acerca de una población pueden ser examinadas a través de información proporcionada por una muestra. Un paso esencial en este proceso es asegurar que tomamos una muestra representativa de la población a analizar.

El uso de las tablas de contingencias es de mayor importancia cuando en un estudio estadístico o no, se requieren probar las hipótesis que generalmente se consideran bajo el tema, y que son; la independencia de las variables o la identidad de las distribuciones poblacionales. Por ejemplo, en la siguiente tabla se puede observar una muestra de 5,375 muertes a causa de tuberculosis

clasificadas con respecto a dos variables cualitativas, llamadas sexo y tipo de causa de muerte por tuberculosis. (Hay que hacer notar que las categorías de las variables son exhaustivas y mutuamente excluyentes.)

Una tabla como ésta, es conocida como una tabla de contingencias, y es un ejemplo de 2x2 que es el caso más simple:

	Hombres	**Mujeres**	**Total**
Tuberculosis del sistema respiratorio.	3,534	1,319	4,853
Otras formas de tuberculosis.	270	252	522
Tuberculosis (todas las formas)	3,804	1,571	5,375

Las entradas en las celdas para estos datos son las frecuencias. Éstas pueden ser transformadas en proporciones o porcentajes, pero es importante notar que, de cualquier forma que sean presentadas, los datos fueron originalmente frecuencias o puntajes antes que medidas. Por supuesto, datos continuos pueden a menudo expresarse de forma discreta utilizando intervalos en la escala continua. La edad, por ejemplo, proporciona datos de este tipo, si se clasificara en grupos, los intervalos correspondientes a estos pueden ser tratados como si ellos fueran unidades discretas.

La tabla del ejemplo anterior comprende sólo dos variables y puede referirse a ella como una tabla de contingencias de dos dimensiones, pero también existen

de tres, cuatro y multidimensionales, cuando la población es clasificada con respecto a más de dos variables cualitativas.

Formalmente una tabla de contingencias puede mostrarse como una clasificación cruzada de los posibles valores (o categorías) de dos o más variables, juntas con un número de observaciones en cada celda de la clasificación cruzada reportada.

3.2 Distribución de frecuencias

Frecuencias

Frecuencia es el número de veces que se repite una observación de un determinado fenómeno.

Tipos de frecuencias:

a) Frecuencias sin acumular. Estas pueden ser:

Frecuencias absolutas (n_i): Se denomina la frecuencia absoluta del nivel i-ésimo de un factor (suponiendo que existen i niveles), al número de veces que el mismo se presenta para los individuos considerados.

Frecuencias relativas (f_i): Frecuencia relativa del nivel i-ésimo, es la relación por cociente entre el número de veces que aparece tal nivel y el número total de los elementos observados; es decir, es el cociente entre la frecuencia absoluta y el total de datos (N).

b) Frecuencias acumuladas

Éstas también pueden ser absolutas o relativas:

Frecuencia absoluta acumulada (N_i): Se define como el número de elementos cuyo nivel es igual o inferior al i-ésimo.

Frecuencia relativa acumulada (F_i): Es la frecuencia absoluta acumulada dividida por el número total de elementos.

3.3 Tablas de contingencias

Considérese una población o muestra, compuesta por N individuos sobre los que se pretende estudiar simultáneamente dos atributos o factores. Se designará por $A_1,...,A_r$ y $B_1,...,B_c$ las r y c modalidades del factor 1 y del factor 2 respectivamente y por n_{ij} el número de individuos que presentan a la vez las modalidades A_i y B_j. La tabla estadística que describe será una tabla de doble entrada, como la siguiente:

		Factor B						
		Nivel 1	Nivel 2	...	Nivel j	...	Nivel c	Total Marginal
Factor A	Nivel 1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.}$
	Nivel 2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Nivel i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i.}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Nivel r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	$n_{r.}$
	Total Marginal	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	N

Se obtiene: $n_{i.} = \sum_{j=1}^{c} n_{ij}$ $n_{.j} = \sum_{i=1}^{r} n_{ij}$

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

Las distribuciones marginales están dadas por:

Factor A	**Frecuencias**	**Factor B**	**Frecuencias**
Nivel 1	$n_{1.}$	*Nivel 1*	$n_{.1}$
Nivel 2	$n_{2.}$	*Nivel 2*	$n_{.2}$
.	.	.	.
.	.	.	.
.	.	.	.
Nivel j	$n_{i.}$	*Nivel j*	$n_{.j}$
.	.	.	.
.	.	.	.
.	.	.	.
Nivel r	$\frac{n_{r.}}{N}$	*Nivel c*	$\frac{n_{.c}}{N}$

y las distribuciones condicionadas:

Factor A/B$_j$	**Frecuencias**	**Factor B/A$_i$**	**Frecuencias**
Nivel 1	n_{1j}	*Nivel 1*	n_{i1}
Nivel 2	n_{2j}	*Nivel 2*	n_{i2}
.	.	.	.
.	.	.	.
.	.	.	.
Nivel j	n_{ij}	*Nivel j*	n_{ij}
.	.	.	.
.	.	.	.
.	.	.	.
Nivel r	$\frac{n_{rj}}{n_{.j}}$	*Nivel c*	$\frac{n_{ic}}{n_{i.}}$

Un caso particular de las tablas de contingencias es el correspondiente a las tablas de 2×2, es decir, en las que los dos factores considerados presentan, cada uno de ellos, únicamente dos categorías mutuamente excluyentes.

3.4 Tipos de tablas de contingencias

Goodman, (1981) listó tres tipos ideales de tablas de contingencias para dos factores:

1.- Una distribución mixta de dos variables explicativas (por ejemplo, estatura y peso).

2.- La relación casual del suceso de una variable dependiente sobre una variable explicativa, (por ejemplo: fumar y cáncer de pulmón).

3.- La asociación entre el resultado de dos variables (por ejemplo, la actitud hacia el aborto y la actitud hacia el sexo prematrimonial).

Es importante notar que la diferencia entre estos tres tipo de tablas de contingencias es conceptual, ya que aparecen en la misma forma.

3.5 Distribuciones de más de dos dimensiones

Suponiendo que se está interesado en analizar conjuntamente tres factores, que se denominarán A, B y C, con r, c y s categorías (niveles o modalidades), respectivamente; la presentación de una población o una muestra cualquiera en forma de tabla se hará de la siguiente manera:

		Factor C: Categoría 1 Factor B				
		1	2	...	c	$n_{i.1}$
Factor A	1	n_{111}	n_{121}	...	n_{1c1}	$n_{1.1}$
	2	n_{211}	n_{221}	...	n_{2c1}	$n_{1.2}$
	⋮	⋮	⋮	⋮	⋮	⋮
	r	n_{r11}	n_{r21}	...	n_{rc1}	$n_{r.1}$
	$n_{.j1}$	$n_{.11}$	$n_{.21}$	...	$n_{.c1}$	$n_{..1}$
		Factor C: Categoría 2 Factor B				
		1	2	...	c	$n_{i.2}$
Factor A	1	n_{112}	n_{122}	...	n_{1c2}	$n_{1.2}$
	2	n_{212}	n_{222}	...	n_{2c2}	$n_{2.2}$
	⋮	⋮	⋮	⋮	⋮	⋮
	r	n_{r12}	n_{r22}	...	n_{rc2}	$n_{r.2}$
	$n_{.j2}$	$n_{.12}$	$n_{.22}$	...	$n_{.c2}$	$n_{..2}$
		Factor C: Categoría s Factor B				
		1	2	...	c	$n_{i.s}$
Factor A	1	n_{11s}	n_{12s}	...	n_{1cs}	$n_{1.s}$
	2	n_{21s}	n_{22s}	...	n_{2cs}	$n_{2.s}$
	⋮	⋮	⋮	⋮	⋮	⋮
	r	n_{r1s}	n_{r2s}	...	n_{rcs}	$n_{r.s}$
	$n_{.js}$	$n_{.1s}$	$n_{.2s}$	...	$n_{.cs}$	$n_{..s}$

En tablas o presentaciones como la anterior se encuentran varios tipos de frecuencias:

- Frecuencias conjuntas tridimensionales, n_{ijk}, que indican el número de individuos que presentan simultáneamente los niveles ijk de los tres factores considerados.
- Frecuencias marginales bivariantes, $n_{ij.}$, $n_{i.k}$ y $n_{.jk}$, definidas como:

$$n_{ij.} = \sum_{k=1}^{s} n_{ijk} \qquad n_{i.k} = \sum_{j=1}^{c} n_{ijk} \qquad n_{.jk} = \sum_{i=1}^{r} n_{ijk}$$

con $i = 1,...,r$, $j = 1,...,c$, y $k = 1,..,s$, y donde $n_{ij.}$ representa el número de individuos que poseen simultáneamente los caracteres i, j de los dos

primeros factores, cualquiera que sea la categoría que les corresponda en el tercero.

- Frecuencias marginales unidimensionales, $n_{i..}$, $n_{.j.}$ y $n_{..k}$, se calculan como:

$$n_{i..} = \sum_{j=1}^{c}\sum_{k=1}^{s} n_{ijk} \qquad n_{.j.} = \sum_{i=1}^{r}\sum_{k=1}^{s} n_{ijk} \qquad n_{..k} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ijk}$$

De lo anterior,

$$N = \sum_{i=1}^{r} n_{i..} = \sum_{j=1}^{c} n_{.j.} = \sum_{k=1}^{s} n_{..k} = \sum_{i,j,k} n_{ijk} \ ,$$

de igual forma:

$$N = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij.} = \sum_{i=1}^{r}\sum_{k=1}^{s} n_{i.k} = \sum_{j=1}^{c}\sum_{k=1}^{s} n_{.jk} = \sum_{i,j,k} n_{ijk} \ ,$$

- Frecuencias condicionadas: el calculo de este tipo de frecuencias es posible considerando fijar condiciones sobre uno de los factores, lo que daría lugar a tres tipos de frecuencias.

$$n_{ij|k} \ , \ n_{i\,k|j} \ \text{y} \ n_{j\,k|i} \ ,$$

o sobre dos de los factores, resultando:

$$n_{i|jk},\ n_{j|ik}\ \text{y}\ n_{k|ij}.$$

3.6 Probabilidad condicional e independencia

Para el análisis de tablas de contingencias, existen dos bases fundamentales. Una de ellas es la definición y uso de los *momios* y dentro de esta misma, la definición y uso del *cociente de momios*. Otra base para el análisis de las tablas de contingencias es el uso de la *independencia* y la *independencia condicional* en las tablas de probabilidades caracterizadas.

Los momios son muy familiares por su uso en eventos deportivos. Éstos son confundidos frecuentemente con probabilidades. En el análisis de modelos log-lineales, los momios y el cociente de momios son usados extensamente.

Suponga un evento, por ejemplo: el día de mañana estará nublado y que tiene una probabilidad *p* de ocurrir, el momio del evento es:

$$momio = \frac{p}{1-p} = \frac{(evento\ ocurra)}{(evento\ no\ ocurra)}.$$

Es decir, suponiendo que la probabilidad de que el día de mañana estará nublado fuese de 0.8, el momio de que el día de mañana estará nublado sería $\frac{0.8}{0.2} = 4$. O sea 4 como $\frac{4}{1}$ lo que se puede interpretar como el momio de que el día de

mañana estará nublado es 4 a 1. El hecho de que el momio sea mayor que 1, indica que el evento tiene una probabilidad de ocurrir mayor a un medio, recíprocamente, si el momio es menor que uno, la probabilidad de que el evento ocurra es menor a un medio.

Mientras más grandes sean los momios, mayor es la probabilidad, mientras más cerca estén de cero, la probabilidad va disminuyendo. De hecho, para probabilidades y momios que están muy cerca del cero, no hay diferencia esencialmente entre los números, por otro lado, cuando las probabilidades se acercan a 1, los momios correspondientes se acercan a infinito.

Dados los momios de que un evento ocurra, la probabilidad del evento es muy fácil de obtener. Si el momio es θ, entonces la probabilidad p, se puede obtener como:

$$p = \frac{\theta}{\theta + 1}$$

Examinando, probabilidades entre cero y un medio, corresponden a momios entre cero y uno:

$$0 < p < \frac{1}{2},$$

$$0 < \frac{\theta}{\theta + 1} < \frac{1}{2},$$

$$0 < \theta < \frac{1}{2}(\vartheta + 1),$$

$$0<\frac{\theta}{2}+\frac{1}{2},$$

$$\frac{\theta}{2}<\frac{1}{2},$$

$$0<\theta<1.$$

Probabilidades entre un medio y uno corresponden a momios entre 1 e infinito.

$$\frac{1}{2}<P<1,$$

$$\frac{1}{2}<\frac{\theta}{\theta+1}<1,$$

$$\frac{\theta+1}{2}<\theta<\theta+1,$$

$$\frac{1}{2}<\frac{\theta}{2} \qquad \theta<1+\theta,$$

$$1<\theta \qquad 0<1,$$

$$\Rightarrow \theta\in(1,\infty),$$

Otra ventaja es el logaritmo de los momios. Probabilidades entre cero y un medio corresponden al logaritmo de momios entre menos infinito y cero.

$$\text{Como } 0<\theta<1 \text{ entonces} -\infty<In\theta<0.$$

Las probabilidades entre un medio y 1 corresponden al logaritmo de momios entre cero e infinito.

$$\theta\in(1,\infty)\Rightarrow 0<In\theta<\infty.$$

La escala de logaritmo de momios es simétrica con respecto al cero como las probabilidades lo son con respecto a un medio. Una unidad por arriba del cero es comparable con una unidad por abajo del mismo.

En el ejemplo anterior, el logaritmo de momios de que el día de mañana estará nublado es log(4), mientras que el logaritmo de momios de que el día de mañana no estará nublado es log(1/4)=-log(4), y se observa que estos números tienen la misma distancia al centro (el cero). Esta simetría en la escala falla en los momios. Los momios de 4 están tres unidades arriba del 1, mientras que los momios de 1/4 están 3/4 por debajo del 1. Para muchos propósitos matemáticos, el logaritmo de momios es una mejor transformación que solamente los momios.

No sólo los momios, sino también el cociente de momios se ofrece naturalmente para el análisis de modelos log-lineales. Así que se vuelve importante desarrollar cierta familiaridad con el cociente de momios.

En el análisis de modelos log-lineales, uno de los usos más comunes del cociente de momios es para observar la igualdad entre conjuntos. Es decir, si el cociente de momios es uno, los dos conjuntos de momios son iguales. Este es un interesante estudio comparativo para decir que los momios de dos cosas son los mismos.

Otro uso común del cociente de momios es para observar que dos de ellos son iguales.

Capítulo 4

Modelos Log-Lineales

En este capítulo se revisará la teoría correspondiente a modelos log-lineales para tablas de dos dimensiones y su generalización a tablas de C×R, revisando sus características principales como lo es independencia y la prueba ji- cuadrada de Pearson, la teoría de máxima verosimilitud para los estimadores y los modelos log-lineales para tablas de dos dimensiones. Para el desarrollo de estos temas se consideraron los libros *Log-Linear Models* y Estadística, Modelos y Métodos.

4.1 Tablas de 2x2

El caso de dos distribuciones binomiales independientes

Se consideran dos binomiales independientes en una tabla de 2x2:

		Factor 2		
		Categoría 1	Categoría 2	Total
Factor 1	Categoría 1	x_{11}	x_{12}	$x_{1.} = x_{11} + x_{12}$
	Categoría 2	x_{21}	x_{22}	$x_{2.} = x_{21} + x_{22}$
	Total	$x_{.1} = x_{11} + x_{21}$	$x_{.2} = x_{12} + x_{22}$	$x_{..} = x_{11} + x_{12} + x_{21} + x_{22}$

Para una tabla de 2x2, los valores observados serán x_{ij}, i=1,2 y j=1,2. Los totales marginales se denotarán $x_{i.} = x_{i1} + x_{i2}$ y $x_{.j} = x_{1j} + x_{2j}$; el total de las observaciones será $x_{..} = x_{11} + x_{12} + x_{21} + x_{22}$. La probabilidad de que una observación se encuentre en el i-ésimo renglón y la j-ésima columna de la tabla está

denotada por p_{ij}. El número esperado de observaciones en el i-ésimo renglón y la j-ésima columna (basado en un modelo estadístico) es denotado por m_{ij}. Para renglones binomiales independientes, se tiene que $m_{ij} = x_{i\cdot} p_{ij}$, para probabilidades y valores esperados son $p_{i\cdot}$ y $p_{\cdot j}$, $m_{i\cdot}$ y $m_{\cdot j}$ correspondientemente. Los totales marginales son definidos como $x_{i\cdot}$.

Cociente de momios en tablas de 2x2

Una técnica muy usada para el análisis de frecuencias de datos es examinar el cociente de momios, el cual está dado por:

$$\frac{\left(\frac{p_{11}}{p_{12}}\right)}{\left(\frac{p_{21}}{p_{22}}\right)} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Si dos binomiales son idénticas, entonces: $p_{11} = p_{21}$ y $p_{12} = p_{22}$.

Lo que implica que:

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = 1.$$

Una alternativa, es utilizar la ji-cuadrada de Pearson, para examinar que cualquier par de binomiales sean iguales. También puede utilizarse el cociente de momios estimado.

Usando $p_{ij} = \frac{x_{ij}}{x_i}$ se tiene que la estimación del cociente de momios está dada por:

$$\frac{p_{11}\ p_{22}}{p_{12}\ p_{21}} = \frac{\frac{x_{11}}{x_1}\ \frac{x_{22}}{x_2}}{\frac{x_{12}}{x_1}\ \frac{x_{21}}{x_2}} = \frac{x_{11}x_{22}}{x_{12}x_{21}}.$$

4.2 Prueba de independencia para una tabla de 2x2

La existencia de factores de respuesta está fuertemente vinculada con el muestreo. Los muestreos de producto de multinomiales son usados comúnmente con una muestra independiente multinomial tomada para toda combinación de factores explicativos y las categorías de multinomiales comenzando con las categorías de los factores explicativos.

En general, las categorías de un factor de respuesta se pueden clasificar con otros o con factores explicativos para producir las categorías en una serie de multinomiales independientes. Algunos factores pueden ser clasificados para definir poblaciones multinomiales mientras que otros como factores de respuesta para definir la categoría de las multinomiales.

Considerando una población dividida en dos categorías y a su vez, cada una de éstas dividida en dos niveles. El primer interés es determinar cuando los renglones son independientes de las columnas y cuando no son independientes.

La probabilidad de que una observación se encuentre en el i-ésimo renglón y la j-ésima columna de la tabla es p_{ij}. La probabilidad de que la observación se encuentre en el i-ésimo renglón es $p_{i.}$. La probabilidad de que la observación se encuentre en la j-ésima columna es $p_{.j}$. Columnas y renglones son independientes si y sólo si para toda $i_{,}$ j se cumple que:

$$p_{ij} = p_{i.} p_{.j} . \qquad (1)$$

El total de las observaciones es n, los puntajes esperados son:

$$m_{ij} = n\, p_{ij} .$$

Si los renglones y columnas son independientes, esto es:

$$m_{ij} = n\, p_{i.} p_{.j} \quad . \qquad (2)$$

No es difícil observar que la condición (1) para la independencia permite considerar que (2) es equivalente a:

$$m_{ij} = \frac{m_{i.} m_{.j}}{n} \quad . \qquad (3)$$

Utilizando la ecuación (1)

$$p_{ij} = p_{i} p_{j} .$$

De la ecuación (2) es fácil ver que:

$$\frac{m_{ij}}{n_{..}} = p_{i} p_{j} .$$

Considerando que renglones y columnas son independientes:

$E(p_{ij}) = E(p_{i} p_{j}) = E(p_{i.})E(p_{.j}) = m_{ij} = m_{i} m_{j}$, entonces:

$$\frac{m_{i} m_{j}}{n_{..}} = p_{i} p_{j} .$$

La estadística de prueba ji-cuadrada de Pearson para probar independencia es:

$$X^{2} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(n_{ij} - m_{ij}^{(0)} \right)^{2}}{m_{ij}^{(0)}} .$$

Donde $m_{ij}^{(0)}$ es una estimación de m_{ij}, basada en el supuesto de que renglones y columnas son independientes. Se toma $m_{i.} = n_{i.}$ y $m_{.j} = n_{.j}$, entonces la ecuación (3) se convierte en:

$$m_{ij}^{(0)} = \frac{n_{i.} n_{.j}}{n} \qquad (4)$$

Se puede llegar a la ecuación (4) a través de (2).

Una estimación obvia de $p_{i.}$ es:

$$p_{i.} = \frac{n_{i.}}{n},$$

similarmente:

$$p_{.j} = \frac{n_{.j}}{n}.$$

Sustituyendo en la ecuación (2), se obtiene la (4).

Los residuales de Pearson están definidos como:

$$r_{ij} = \frac{n_{ij} - \hat{m}_{ij}^{(0)}}{\sqrt{\hat{m}_{ij}^{(0)}}}.$$

En análisis de regresión estándar, es una práctica común el utilizar los residuales para verificar que los supuestos hechos en el modelo son válidos y poder detectar la presencia de observaciones que usualmente intervienen en el ajuste de éste. Para modelos log-lineales hay una influencia análoga que depende del diseño y de las probabilidades de que las observaciones caigan en una celda en particular. Como las probabilidades no son conocidas, tienen que ser estimadas, entonces se usan *influencias estimadas* las cuales en el *i-ésimo* caso se denotan como:

$$\hat{a}_{ii} .$$

El modelo log-lineal análogo del *residual estándarizado* es:

$$r_i = \frac{n_i - m\hat{n}_i}{\sqrt{m\hat{n}_i(1-\hat{a}_{ii})}} .$$

Para un modelo correcto y una muestra grande, una aproximación para la distribución de r_i es una normal estándar.

Cociente de momios (Independencia)

El cociente de momios puede ser utilizado para analizar la independencia de dos factores en una prueba multinomial.

Proposición.

Si renglones y columnas son independientes, entonces el cociente de momios es igual a 1.

$$\frac{\left(\frac{p_{11}}{p_{21}}\right)}{\left(\frac{p_{12}}{p_{22}}\right)} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = 1 .$$

Demostración:

Por la ecuación $p_{ij} = p_{i.} p_{.j}$:

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{p_{1.}p_{.1}p_{2.}p_{.2}}{p_{1.}p_{.2}p_{2.}p_{.1}} = 1 .$$

Si el cociente de momios es estimado bajo el supuesto de independencia,

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j} = \frac{n_{i.}n_{.j}}{(n)^2} .$$

De este modo, el cociente de momios estimado es siempre 1. Una aproximación interesante del cociente de momios estimado es, sin suponer independencia, observar como el cociente se acerca a 1.

Con esta aproximación $\hat{p}_{ij} = \frac{n_{ij}}{n}$ y $\frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$.

4.3 Tablas de C×R

Las situaciones analizadas anteriormente pueden ser generalizadas considerando muestras de C poblaciones diferentes, que se originan de una distribución multinomial, cada una de las cuales se divide en R categorías. Esto es, una muestra de producto de multinomiales. Asumiendo que un factor tenga C y el otro tenga R categorías, entre los dos la población es dividida dentro del total en $C \times R$ categorías. La distribución de los puntajes dentro de las $C \times R$ categorías es asumida como una distribución multinomial. Por consiguiente, este muestreo es llamado muestreo multinomial.

En una tabla de $C \times R$ con observaciones n_{cr}, $c=1,\ldots,C$, $r=1,\ldots,R$, puede escribirse como:

Factor 2

(Categorías)

Factor 1

(Poblaciones)

n_{cr}	1	2	...	R	Total
1	n_{11}	n_{12}	...	n_{1R}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2R}	$n_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
C	n_{C1}	n_{C2}	...	n_{CR}	$n_{C\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot R}$	$n_{\cdot\cdot}$

Se puede representar con matrices similares a las probabilidades p_{ij} y los valores esperados m_{ij}.

El análisis del muestreo de producto de multinomiales se comienza probando si todas las *C* poblaciones multinomiales son idénticas. En otras palabras la prueba:

$$H_0 : p_{1r} = p_{2r} = \ldots = p_{Cr} \text{ para toda } r=1,2,\ldots,R. \qquad (1)$$

vs

H_A : El modelo (1) es falso.

Esta prueba es llamada de homogeneidad de proporciones.

Se utiliza la estadística de prueba ji-cuadrada de Pearson para evaluar la aptitud de la hipótesis nula del modelo, para lo cual se requiere estimar los valores esperados m_{cr}. Si cada muestra tiene una distribución multinomial, entonces:

$$m_{cr} = n_{c\cdot} p_{cr}.$$

Si H_0 es verdadera, p_{cr} es la misma para todos los valores de *c*. Una estimación común del valor de las p_{cr} es:

$$\hat{p}_{cr}^{(0)} = \frac{n_{\cdot r}}{n}.$$

De este hecho se obtiene:

$$\hat{m}_{cr}^{(0)} = n \left(\begin{matrix} n_{+r} \\ n \end{matrix} \right).$$

En ambas, $\hat{p}_{cr}^{(0)}$ y $\hat{m}_{cr}^{(0)}$, el índice (0) es utilizado para indicar que la estimación fue obtenida bajo el supuesto de que H_0 es verdadera. La estadística de prueba de ji-cuadrada de Pearson es:

$$\chi^2 = \sum_{c=1}^{C} \sum_{r=1}^{R} \frac{\left(n_{cr} - \hat{m}_{cr}^{(0)}\right)^2}{\hat{m}_{cr}^{(0)}}.$$

Para muestras grandes, si H_0 es verdadera, la aproximación:

$$\chi^2 \sim \chi^2_{((C-1)(R-1))}.$$

Es válida. H_0 se rechaza con un nivel de significancia α en la prueba si:

$$\chi^2 > \chi^2_{(1-\alpha,(C-1)(R-1))}.$$

El análisis de una muestra multinomial comienza probando la independencia de los dos factores. En particular, si se quiere probar el modelo:

$$H_0 : p_{cr} = p_{c+} \times p_{+r} \quad c = 1,\ldots,C, \; r = 1,\ldots,R \qquad (2)$$

Se utilizará de nuevo la ji-cuadrada de Pearson. Las probabilidades marginales son estimadas como:

$$\hat{p}_{c.} = \frac{n_{c.}}{n_{..}}$$

y

$$\hat{p}_{.r} = \frac{n_{.r}}{n_{..}}.$$

Porque $m_{cr} = n_{..}p_{cr}$. Si el modelo (2) es verdadero, se puede estimar a m_{cr} con:

$$\hat{m}_{cr}^{(0)} = n_{..}\hat{p}_{c.}\hat{p}_{.r},$$

$$= n_{..}\left(\frac{n_{c.}}{n_{..}}\right)\left(\frac{n_{.r}}{n_{..}}\right),$$

$$= \frac{n_{c.}n_{.r}}{n_{..}}.$$

Donde el índice (0) en $\hat{m}_{cr}^{(0)}$ indica que la estimación es obtenida suponiendo que (2) es cierta. Además, si (2) es verdadera y la muestra es muy grande, se distribuye aproximadamente como $\chi^2_{((C-1)(R-1))}$. H_0 se rechaza con un nivel de significancia α en la prueba si:

$$\chi^2 > x^2_{(1-\alpha,(C-1)(R-1))}.$$

4.4 Teoría de máxima verosimilitud para tablas de dos dimensiones.

Para trabajar la teoría de máxima verosimilitud, primero se introducirá el siguiente lema.

Lema: Sea $f(p_1,\ldots,p_r)=\sum_{i=1}^{r} n_i \log p_i$. Si $n_i>0$ para $i=1,\ldots,r$ entonces, subordinando a las condiciones $0<p_i<1$ y $p_.=1$, el máximo de $f(p_1,\ldots,p_r)$ es obtenido en el punto $(p_1,\ldots,p_r)=(\hat{p}_1,\ldots,\hat{p}_r)$ donde $\hat{p}_i=\frac{n_i}{n}$.

Considérese una muestra de producto de multinomiales de I poblaciones, con cada una dividida en J categorías iguales. Las I poblaciones formarán los renglones de una tabla de $I\times J$. Ningún resultado será representado por una muestra multinomial en una tabla de $I\times J$.

La probabilidad de obtener el dato $n_{i1},\ldots,n_{ij}$, de la i-ésima muestra multinomial es:

$$\frac{n_i!}{\prod_{j=1}^{J} n_{ij}!}\prod_{j=1}^{J} p_{ij}^{n_{ij}} .$$

Como las I multinomiales son independientes, la probabilidad de obtener todos los valores n_{ij} , $i=1,\ldots,I$, $j=1,\ldots,J$, es:

$$\prod_{i=1}^{I}\left[\frac{n_{i.}!}{\prod_{j=1}^{J} n_{ij}!}\prod_{j=1}^{J} p_{ij}^{n_{ij}}\right]. \qquad (1)$$

En este caso, si se conocieran las p_{ij}, podría encontrarse la probabilidad de obtener algún conjunto de n_{ij}. De hecho, se está precisamente en la posición opuesta, debido a que no se conocen las p_{ij}, pero se conocen las n_{ij}, ya que éstas han sido observadas. Si se piensa en (1) como la función de las p_{ij}, entonces puede escribirse:

$$L(p) = \prod_{i=1}^{I}\left[\frac{n_{i.}!}{\prod_{j=1}^{J} n_{ij}!}\prod_{j=1}^{J} p_{ij}^{n_{ij}}\right] \qquad (2)$$

donde $p = (p_{11}, p_{12}, \ldots, p_{IJ})$. $L(p)$ es llamada la función de verosimilitud para *p*.

Algunos valores de *p* dan una probabilidad pequeña de observar las n_{ij} que son actualmente analizadas. Es improbable que dichos valores de *p* sean los verdaderos ya que el verdadero es probablemente algún valor que da una relativa gran probabilidad de contemplar lo que se estaba realmente observando. Si se desea estimar *p*, tiene sentido utilizar un valor de *p* que de una gran probabilidad de ver lo que se estaba examinando. En otras palabras, tiene sentido estimar *p* con un valor $\hat{p}$ que maximice la función de verosimilitud *L(p)* dicho valor es llamado estimador de máxima verosimilitud de *p* (***MLE***).

La función de máxima verosimilitud, se obtiene al maximizar el logaritmo de la función de verosimilitud, en el cual los productos cambian a sumas. Como el logaritmo es estrictamente una función creciente, el máximo de la verosimilitud y el máximo del log de la verosimilitud ocurren en el mismo punto.

Para el producto de una muestra de multinomiales, la función log-verosimilitud es:

$$\log L(p) = \sum_{i=1}^{I} \left[\log(n_i!) - \sum_{j=1}^{J} \log(n_{ij}!) + \sum_{j=1}^{J} n_{ij} \log p_{ij} \right].$$

Al maximizar ésta como una función de las p_{ij}, se pueden ignorar algunos términos que no dependen de p_{ij}. Es suficiente maximizar:

$$l(p) = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log p_{ij} .$$

El máximo se encuentra cuando se maximiza cada uno de los términos $\sum_{j=1}^{J} n_{ij} \log p_{ij}$. Por el lema mencionado al inicio de este tema, el máximo es encontrado en $p = \hat{p}$, donde:

$$\hat{p} \equiv \frac{n_{ij}}{n_i} .$$

También puede obtenerse el estimador de máxima verosimilitud por los puntajes esperados de m_{ij}, porque $m_{ij} = n_i \hat{p}_{ij} = n_{ij}$.

Esto se deriva de la invarianza de los estimadores de máxima verosimilitud; para algún parámetro θ y $\hat{\theta}$, la estimación de máxima verosimilitud de una función θ, es decir $f(\theta)$, es la correspondiente función de máxima verosimilitud de la , $f(\hat{\theta})$.

Si se cambia el modelo, entonces la hipótesis nula se transforma en:

$$H_0 : p_{1j} = \ldots = p_{Ij} \text{ , } j=1,\ldots,J$$

Es verdadera si se consiguen diferentes estimadores de máxima verosimilitud. Sea $\pi_j = p_{1j} = \ldots = p_{Ij}$. La función log-verosimilitud será:

$$\log L(p) = \sum_{i=1}^{I}\left[\log(n_{i.}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log\pi_j\right].$$

Considerando que pueden ignorarse los términos que no envuelven las p_{ij}, tiene que maximizarse en primer lugar:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\pi_j .$$

Esto es equivalente a:

$$\sum_{j=1}^{J} n_{ij} \log \pi_j .$$

Por el lema mencionado anteriormente, los estimadores de máxima verosimilitud serán:

$$\hat{p}_{ij}^{(0)} = \hat{\pi}_j = \frac{n_{ij}}{n} .$$

donde el (0) es usado en $\hat{p}_{ij}^{(0)}$ para indicar que la estimación es obtenida asumiendo que H_0 es verdadera.

Los estimadores de máxima verosimilitud de las m_{ij} no son difíciles de obtener bajo el modelo nulo H_0, como $\hat{m}_{ij} = n_i \hat{p}_{ij}$, entonces:

$$\hat{m}_{ij}^{(0)} = n_i \hat{p}_{ij}^{(0)} = \frac{n_i n_{ij}}{n} .$$

Obsérvese que $\hat{p}_{ij}^{(0)}$ y $\hat{m}_{ij}^{(0)}$ son precisamente los estimadores utilizados en el tema de tablas de C×R para la prueba de H_0.

La función de máxima verosimilitud puede usarse también como una prueba básica de modo que H_0 es verdadera. Los datos tienen cierta verosimilitud de ser observados y pueden ser resumidos como el valor máximo que la función de verosimilitud alcanza. Si se pone alguna restricción en los valores posibles de las

p_{ij}, se reduce la verosimilitud de los valores observados. Si al ponerse restricciones sobre las p_{ij} se reduce en gran parte la verosimilitud, puede inferirse que las restricciones sobre las p_{ij} no son idóneas para ser válidas. La relativa reducción en la verosimilitud puede ser medida observando al máximo de $L(p)$ sometido a la restricción dividida por la longitud total del máximo de $L(p)$. Si la proporción obtenida es muy pequeña, se rechazará que los supuestos que restringen a las p_{ij} son válidos. En particular, si las restricciones sobre las p_{ij} son que H_0 es verdadera, se rechazará H_0 cuando la proporción de verosimilitud es muy pequeña.

Si se simplifican las matemáticas examinando el logaritmo de la proporción de verosimilitud y rechazando H_0 cuando el log se hace muy pequeño. Por supuesto el log de la proporción de verosimilitud es justamente la diferencia entre el log-verosimilitud. El valor máximo del log-verosimilitud cuando se reduce el modelo H_0 es verdadero si:

$$\log L(\hat{p}^{(0)}) = \sum_{i=1}^{I}\left[\log(n_i!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{.j}}{n_{..}}\right)\right].$$

La longitud total del máximo del log-verosimilitud es:

$$\log L(\hat{p}) = \sum_{i=1}^{I}\left[\log(n_i!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{n}\right)\right].$$

La diferencia es:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{.j}}{n_{..}}\right) - \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{n_{i.}}\right).$$

Si se multiplica por -2 y se simplifica, se obtiene una prueba estadística para la proporción de verosimilitud:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{m}_{ij} \log\left(\frac{\hat{m}_{ij}}{\hat{m}_{ij}^{(0)}}\right).$$

donde $\hat{m}_{ij} = n_{ij}$ es el estimador de máxima verosimilitud (EMV) de m_{ij} en el modelo no restringido y $\hat{m}_{ij}^{(0)} = \frac{n_{i.}n_{.j}}{n_{..}}$ es el *EMV* de m_{ij} bajo la restricción de que H_0 es verdadera.

La razón de multiplicar por -2, es que con esta multiplicación, la aproximación:

$$G^2 \sim \chi^2_{((I-1)(J-1))}.$$

Es válida cuando H_0 es verdadera y las muestras son muy grandes. Obsérvese que como H_0 fue rechazada para valores muy pequeños de la proporción de verosimilitud, después se toma el logaritmo y se multiplica por -2, H_0 debería ser rechazada para valores grandes de G^2. En particular, para muestras grandes, un nivel α de H_0 es rechazada sí:

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

4.5 Modelos Log-Lineales para tablas de dos dimensiones

La intención es explotar las similitudes entre el análisis de varianza (ANOVA) y regresión, por un lado y modelos log-lineales por otro. Se iniciará por discutir el análisis de varianza para dos factores.

Considérese el siguiente modelo balanceado ANOVA:

$$u_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \lambda_{ijk}$$

Pueden cambiarse los símbolos utilizados para denotar parámetros, y reescribir el modelo como:

$$y_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} + \lambda_{ijk} \qquad (1)$$

$i = 1.....I, j = 1.....J$ y $k = 1.....K$. Las λ_{ijk} son variables aleatorias que miden el error, asumidas para ser $N(0,\sigma^2)$ independientes. Puede estimarse σ^2 y probarse su interacción. Si existen, puede observarse también el contraste en esta interacción, si no , pueden probarse los efectos importantes y observar los diferencias en éstos. Si algún nivel del factor corresponde a valores cuantitativos, entonces las teorías de regresión pueden ser incorporadas en la ANOVA. La estimación de σ^2 es el error cuadrático medio (EMC):

$$ECM = \frac{1}{IJ(K-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{ijk} - \bar{y}_{ij})^2 .$$

El estimador de σ^2 es una función de las y_{ijk}. En particular, puede formarse una tabla de $I \times J$ de las y_{ijk}. El objetivo del análisis es explorar la estructura de esta tabla. El modelo ANOVA (1) y los correspondientes contrastes en las interacciones y los efectos principales han probado ser herramientas muy útiles para explorar tablas de $I \times J$.

Se considerará que el modelo ANOVA es básicamente efectivo, que las y_{ijk} son independientes y que:

$$y_{ijk} \sim N(m_{ij}, \sigma^2),$$

donde:

$$m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}. \qquad (2)$$

Otro de los objetivos es examinar la estructura de las m_{ij}. Hacer que se usen los $EMV's$ de las m_{ij}, los cuales son:

$$\hat{m}_{ij} = \bar{y}_{ij}.$$

El análisis o planteamiento del estudio está basado en el hecho de que las m_{ij} son independientes con:

$$\hat{m}_{ij} \sim N\left(m_{ij}, \frac{\sigma^2}{k}\right),$$

y que el ECM es una estimación de σ^2, el cual es independiente de las m_{ij}. Es importante hacer notar que aunque el ECM no es el EMV de σ^2, exactamente la misma prueba de intervalos de confianza de las m_{ij} sería obtenida si el EMV para σ^2 fuera usado en lugar del ECM (con un ajuste adecuado en las distribuciones que fueron hechas).

Si se impone la restricción en las m_{ij}, por ejemplo, de no interacción:

$$m_{ij} = u + u_{1(i)} + u_{2(j)}. \qquad (3)$$

Los EMV de las m_{ij} cambiarían. En particular:

$$\hat{m}_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}), \qquad (4)$$

y los EMV de σ^2 también cambian. Esto puede ser usado para mostrar que la prueba usual F para no interacción es justamente la prueba de cociente de verosimilitud para no interacción.

Al examinar una tabla de I×J de contingencias, se utilizan técnicas similares. Las entradas de la tabla tienen la propiedad de:

$$E(n_{ij}) = m_{ij}.$$

Asi se tiene interés en la estructura de las m_{ij}; de cualquier modo, en lugar de considerar modelos lineales como (2) y (3), considérense algunos como:

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

y

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}.$$

El análisis contará con los EMV de las m_{ij} de nuevo, y también con las pruebas de cociente de verosimilitud. De cualquier modo existen algunas diferencias. Las n_{ij} son típicamente multinomiales o producto de multinomiales. Tradicionalmente, las pruebas de intervalos de confianza han sido basadas en una gran muestra de distribuciones aproximadas. Por el otro lado, la distribución multinomial depende sólo de las p_{ij} o equivalentemente las m_{ij}, por eso no es necesario dar con término análogo de σ^2 en la teoría normal.

Finalmente, el modelo ANOVA (1) está balanceado, es decir, se tienen k observaciones en cada celda de la tabla. Este balance lleva a simplificaciones en el análisis. Por ejemplo, la fórmula (4) para el EMV de las m_{ij} bajo el modelo de no interacción no aplica. Los modelos log-lineales son análogos a los modelos ANOVA con un insuficiente número de observaciones. Casi nunca se exponen todas las simplificaciones asociadas con observaciones balanceadas en ANOVA y ocasionalmente sólo tienen fórmulas simples para el EMV de las m_{ij}. Aunque mucho del trabajo en modelos log-lineales ha sido utilizado en grandes muestras (distribuciones asintóticas), recientemente se ha tenido un trabajo considerable en inferencia condicional exacta para pequeñas muestras.

Existen razones para escribir modelos tipo ANOVA para $\log(m_{ij})$ en lugar de las m_{ij}. Una es que la teoría de una muestra grande puede ser resuelta. En otras palabras, la explicación para hacerlo es porque realmente puede hacerse. Otra de las razones es que los modelos log-lineales aparecen de manera natural de las matemáticas de Poisson (muestras), multinomiales que tienen celdas con valores esperados y que están limitadas entre 0 y una muestra de tamaño N. Esto limita los parámetros de los modelos tipo ANOVA para m_{ij}. Dichos problemas no aparecen en los modelos log-lineales, además de que estos ofrecen tener frecuentemente buenas interpretaciones, las cuales se examinarán a continuación para dos factores.

Considérese una muestra multinomial. Se sabe que $m_{ij} = n\, p_{ij}$, puede escribirse el siguiente modelo:

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}. \qquad (5)$$

El cual no tiene ningún término perfecto. Los términos u, $u_{1(i)}$ y $u_{2(j)}$ son totalmente redundantes. Estos pueden tener valores en todo y sin embargo el término $u_{12(ij)}$ puede ser elegido de tal manera que la ecuación (5) puede ser obtenida. Porque el modelo (5) tiene suficientes términos en u para explicar completamente algún conjunto de m_{ij}, el modelo (5) es llamado *saturado*.

Un ejemplo más interesante de un modelo log-lineal ocurre cuando los renglones y columnas de una tabla son independientes entre si:

$$m_{ij} = n\, p_{i+} p_{+j}.$$

Entonces:

$$\log m_{ij} = \log n + \log p_{i.} + \log p_{.j}.$$

En otras palabras, si renglones y columnas son independientes en un modelo log-lineal de la forma:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \qquad (6)$$

es sustentado, de cualquier modo, si se está basado el análisis en modelos log-lineales, es igualmente importante conocer si el modelo (6) tiene entonces renglones y columnas independientes.

Teorema: Para una muestra multinomial en una tabla de $I \times J$, $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$, $i = 1,\ldots,I$, $j = 1,\ldots,J$, si y sólo si $p_{ij} = p_{i.}.p_{.j}$, $i = 1,\ldots,I$, $j = 1,\ldots,J$.

Demostración:

Tiene que demostrarse que la independencia implica un modelo log-lineal.

Si el modelo log-lineal es sustentado, entonces:

$$m_{ij} = e^{u + u_{1(i)} + u_{2(j)}}$$

Sea $a = e^{u}$, $a_{1(i)} = e^{u_{1(i)}}$ y $a_{2(j)} = e^{u_{2(j)}}$. Sea $a_{1(\cdot)} = \sum_{i=1}^{I} a_{1(i)}$ y $a_{2(\cdot)} = \sum_{j=1}^{J} a_{2(j)}$. Obsérvese que:

$$p_{ij} = \frac{m_{ij}}{N} = \frac{aa_{1(i)}a_{2(j)}}{N},$$

$$p_{i.} = \frac{aa_{1(i)}a_{2(.)}}{N},$$

$$p_{.j} = \frac{aa_{1(.)}a_{2(j)}}{N},$$

y $$1 = p_{..} = \frac{aa_{1(.)}a_{2(.)}}{N}.$$

Sustituyendo se obtiene:

$$p_{i.}p_{.j} = \frac{aa_{1(i)}a_{2(.)}aa_{1(.)}a_{2(j)}}{N^2},$$

$$p_{i.}p_{.j} = \left(\frac{aa_{1(i)}a_{2(j)}}{N}\right)\left(\frac{aa_{1(.)}a_{2(.)}}{N}\right),$$

$$p_{i.}p_{.j} = \left(\frac{aa_{1(i)}a_{2(j)}}{N}\right),$$

$$p_{i.}p_{.j} = P_{ij}.$$

De esta manera el modelo log-lineal implica independencia.

Para una muestra de producto de multinomiales:

$$m_{ij} = n\ p_{ij}, \qquad (7)$$

y el modelo log-lineal:

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)},$$

es trivial. Ahora considérese el modelo bajo H_0. Obsérvese que si $\pi_j = p_{1j} = \ldots = p_{Ij}$ para toda $j = 1,\ldots,J$ entonces:

$$m_{ij} = n_i \pi_j$$

Teorema: Para una muestra de producto de multinomiales en una tabla de I×J, donde los renglones son muestras independientes, $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$, $i = 1,\ldots,I$, $j = 1,\ldots,J$, si y sólo si $p_{1j} = \ldots = p_{Ij}$, $j = 1,\ldots,J$.

Demostración:

Si para cada j las probabilidades p_{ij} son iguales, se tiene que $m_{ij} = n_i \pi_j$ y $\log(m_{ij}) = \log n_i + \log \pi_j$. Haciendo $u = 0$, $u_{1(i)} = \log(n_i)$ y $u_{2(j)} = \log(\pi_j)$, se muestra que el modelo log-lineal sustenta.

Recíprocamente, si $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$, entonces $m_{ij} = a a_{1(i)} a_{2(j)}$, donde $a = e^u$, $a_{1(i)} = e^{u_{1(i)}}$ y $a_{2(j)} = e^{u_{2(j)}}$. Nótese que $p_I = 1$, entonces por (7), $m_I = n_I$, y

$$n_i = a a_{1(i)} a_{2(J)}.$$

Porque
$$p_{ij} = \frac{m_{ij}}{n_i},$$

$$= \frac{a a_{1(i)} a_{2(j)}}{a a_{1(i)} a_{2(J)}},$$

$$= \frac{a_{2(j)}}{a_{2()}} .$$

Esto es verdadero para toda i, de esta manera $\frac{a_{2(j)}}{a_{2()}} = p_{1j} = p_{2j} = \ldots = p_{Ij}$, para $j = 1,\ldots,J$.

Cociente de momios

En aplicaciones con tablas de grandes dimensiones, es raro observar que no existan interacciones importantes, por lo que para poder explorar su naturaleza se necesita observar los contrastes entre ellas. Para esto, se necesita un método para definir contrastes en las iteraciones. Se comenzará por revisar los métodos de interacción en el análisis de varianza.

$$\sum_{i=1}^{I}\sum_{j=1}^{J} q_{ij} m_{ij} \qquad (8)$$

Es un contraste en las interacciones. Usando el modelo (2), y el hecho de que $q_{i.} = q_{.j}$, el contraste definido en (8) puede ser escrito también como:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} q_{ij} u_{12(ij)}$$

el cual envuelve sólo las interacciones. La mejor forma de interpretar la obtención del contraste en las interacciones es definirlo en términos de los efectos principales. Sea a_i, $i = 1,\ldots,I$, es un contraste en los renglones (así $a = 0$) y sea b_j, $j = 1,\ldots,J$ al correspondiente en las columnas (entonces $b = 0$). De este modo, si se toma $q_{ij} = a_i b_j$, se obtendrá un contraste en las interacciones.

Es necesario recordar que si no existe interacción, su correspondiente contraste es igual a cero. Recíprocamente, la interacción tiene $(I-1)(J-1)$ grados de libertad, entonces especificando que algún $(I-1)(J-1)$ contraste en la interacción linealmente independiente, sean todos cero, es equivalente especificar que no hay interacción.

Una valiosa técnica analítica para examinar las interacciones en dos direcciones de análisis de varianza es la gráfica. Ésta consiste en representar las I curvas determinadas para conectar los puntos $(j,\hat{m}_{ij})$, $j=1,...,J$, con una línea segmentada. En esta gráfica, $\hat{m}_{ij}=\bar{y}_{ij\cdot}$, la estimación de m_{ij} en el modelo (2). Si no existe interacción, $m_{ij}=u+u_{1(i)}+u_{2(j)}$ y las I curvas teóricas $(j,\hat{m}_{ij})$ son paralelas. Si existe interacción, las I curvas teóricas no son paralelas. Las curvas $(j,\hat{m}_{ij})$, estiman a las teóricas. Si las $(j,\hat{m}_{ij})$ son aproximadamente paralelas, esto sugiere que no existe interacción. Si existe interacción, las curvas estimadas pueden sugerir la naturaleza de ésta. Si las gráficas son aproximadamente paralelas depende de la variabilidad de las $\hat{m}_{ij}$.

Mejor que graficar las I curvas basadas en $(j,\hat{m}_{ij})$, pueden graficarse las J curvas basadas en $(i,\hat{m}_{ij})$, $i=1,...,I$. De nuevo, en ausencia de interacciones, las curvas deberán ser aproximadamente paralelas. Si la manera de tratar las columnas corresponde a niveles cuantitativos, es decir, x_j, $j=1,...,J$, entonces las gráficas de $(x_j,\hat{m}_{ij})$ son apropiadas. Otra vez, se observa paralelismo. Gráficas similares pueden ser construidas si la manera de tratar los renglones es con niveles cuantitativos.

En modelos log-lineales, pueden ser aplicados los mismos procedimientos para $\log(m_{ij})$. En particular, especificando que un cociente de momios igual a uno es

equivalente a especificar que un contraste en la interacción es igual a cero. Primeramente obsérvese que los cocientes de los momios pueden ser escritos en términos de valores esperados. Para una muestra de producto de multinomiales:

$$m_{ij} = n_i p_{ij} .$$

y para una muestra multinomial

$$m_{ij} = n\, p_{ij} .$$

En otro caso,

$$\frac{p_{ij} p_{ij}}{p_{i\cdot} p_{\cdot j}} = \frac{m_{ij} m_{ij}}{m_{i\cdot} m_{\cdot j}} .$$

Si

$$\frac{m_{ij} m_{ij}}{m_{i\cdot} m_{\cdot j}} = 1 ,$$

entonces tomando logaritmos se obtiene

$$\log m_{ij} - \log m_{i\cdot} - \log m_{\cdot j} + \log m_{\cdot\cdot} = 0 .$$

Ésta es precisamente la aseveración de que el contraste de interacción

$$\sum_{r=1}^{I}\sum_{s=1}^{J} q_{rs}\log(m_{rs}). \qquad (9)$$

es igual a cero, donde $q_{11} = q_{1\cdot} = 1$ y $q_{rs} = 0$ para todo par de (r,s). En particular los coeficientes q_{rs} pueden ser obtenidos combinando el contraste en los renglones $a_1 = 1$, $a_i = -1$, $a_r = 0$ para toda r, con el contraste en las columnas $b_1 = 1$, $b_j = -1$, $b_s = 0$, para toda otra s. Obsérvese que el contraste (9) puede también se escrito como:

$$\sum_{r=1}^{I}\sum_{s=1}^{J} q_{rs}u_{12(rs)},$$

donde se ha usado el modelo (5) y el hecho de que $q_{r\cdot} = q_{\cdot s} = 0$.

Si se especifica que:

$$\frac{m_{11}m_{ij}}{m_{1j}m_{i1}} = 1.$$

para toda $i = 2,\ldots,I$ y $j = 2,\ldots,J$, entonces tiene que especificarse que $(I-1)(J-1)$ contrastes de interacciones linealmente independientes en el $\log(m_{ij})$ son todos iguales a cero, en consecuencia no hay interacción.

Al igual que el análisis de varianza, una gráfica de interacción puede ser una valiosa herramienta en el análisis de modelos log-lineales. Las I curvas que conectan las series de puntos $(j,\log(\hat{m}_{ij}))$, $j = 1,\ldots,J$ son la base para la gráfica de interacción. Los valores esperados estimados $\hat{m}_{ij}$ son evaluados usando el modelo (5) con interacción. Bajo el modelo (5), $\hat{m}_{ij} = n_{ij}$. Las I curvas estiman las curvas

teóricas basadas en $(j, \log(\hat{m}_{ij}))$. Si no hay interacción, las curvas teóricas son paralelas y las curvas estimadas deben indicar esto. Si existe interacción, la naturaleza de esta debe sugerirse por las curvas estimadas.

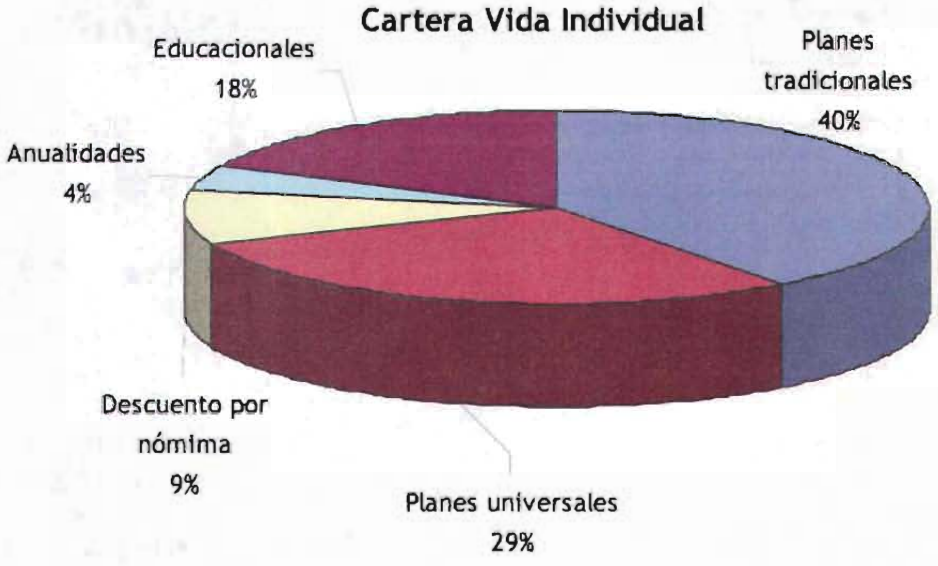
Capítulo 5

Aplicación de un Modelo Log-Lineal a los siniestros de una cartera de seguros de vida individual.

Para esta aplicación, se eligieron los siniestros ocurridos durante un año, de una cartera de seguros de vida individual, en toda la República.

5.1 Introducción

La cartera de expuestos de seguros de vida individual contempla planes tradicionales, universales y flexibles. De lo anterior, se tiene la siguiente distribución:

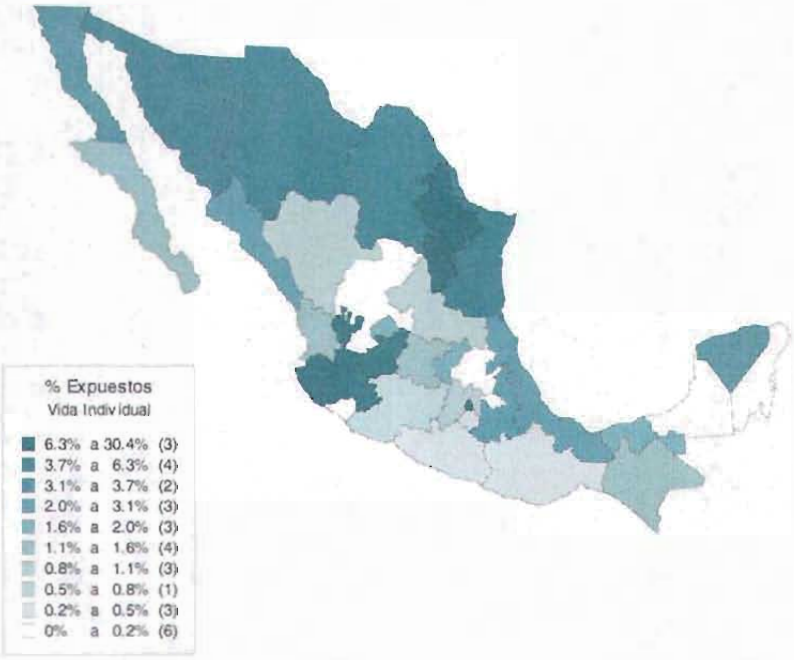


Como puede observarse en la gráfica, los planes tradicionales, que contemplan ordinarios de vida y temporales, son los más frecuentes, y junto con los universales, componen alrededor del 70% de la cartera. Las anualidades

representan el porcentaje de cartera menor, debido a que en México no se tiene una cultura de prever con una pensión adicional a la de IMSS o ISSSTE.

Distribución de expuestos en la República Mexicana

En la República Mexicana, se tiene la siguiente distribución de expuestos por estado.

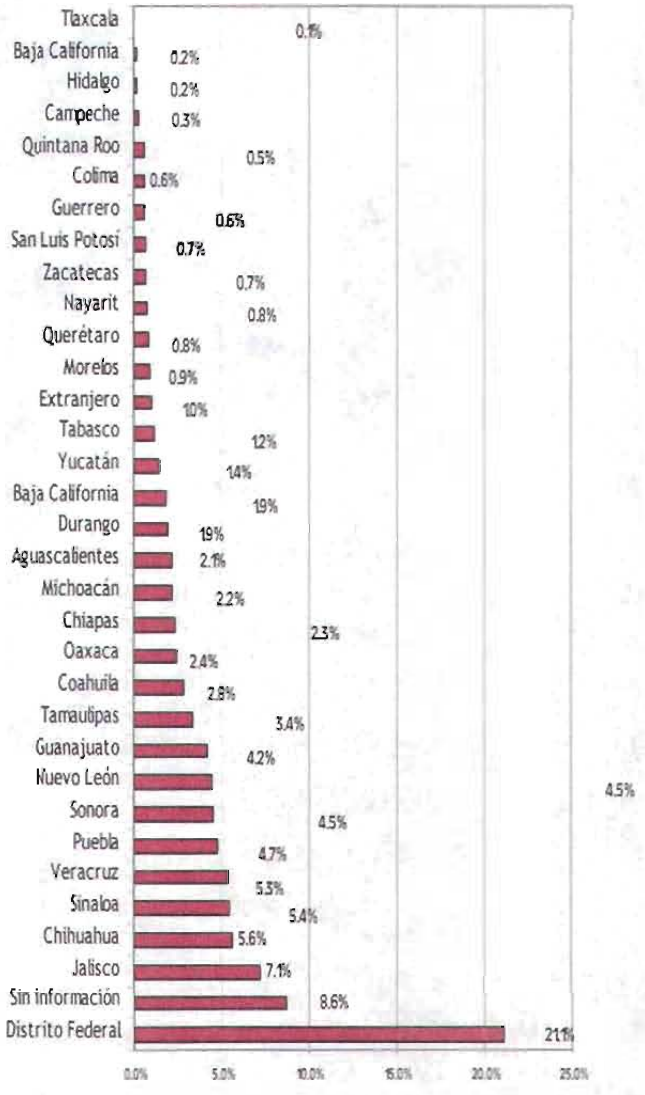


Como puede observarse, en la mayoría de los estados se tiene una participación, aunque en algunos mínima. Los más representativos son el Distrito Federal con 30.4%, Nuevo León con 10.1%, Jalisco con 8.3%, Chihuahua con 6.3% y Sonora con 5.3%, los estados que tiene cero participación son Zacatecas y Tlaxcala.

En las siguientes gráficas se observará la distribución que tuvieron los siniestros de esta cartera, por estado, ocupación, fumador y no fumador, y edad, debido a que son las características que frecuentemente se analizan para la suscripción de estos riesgos.

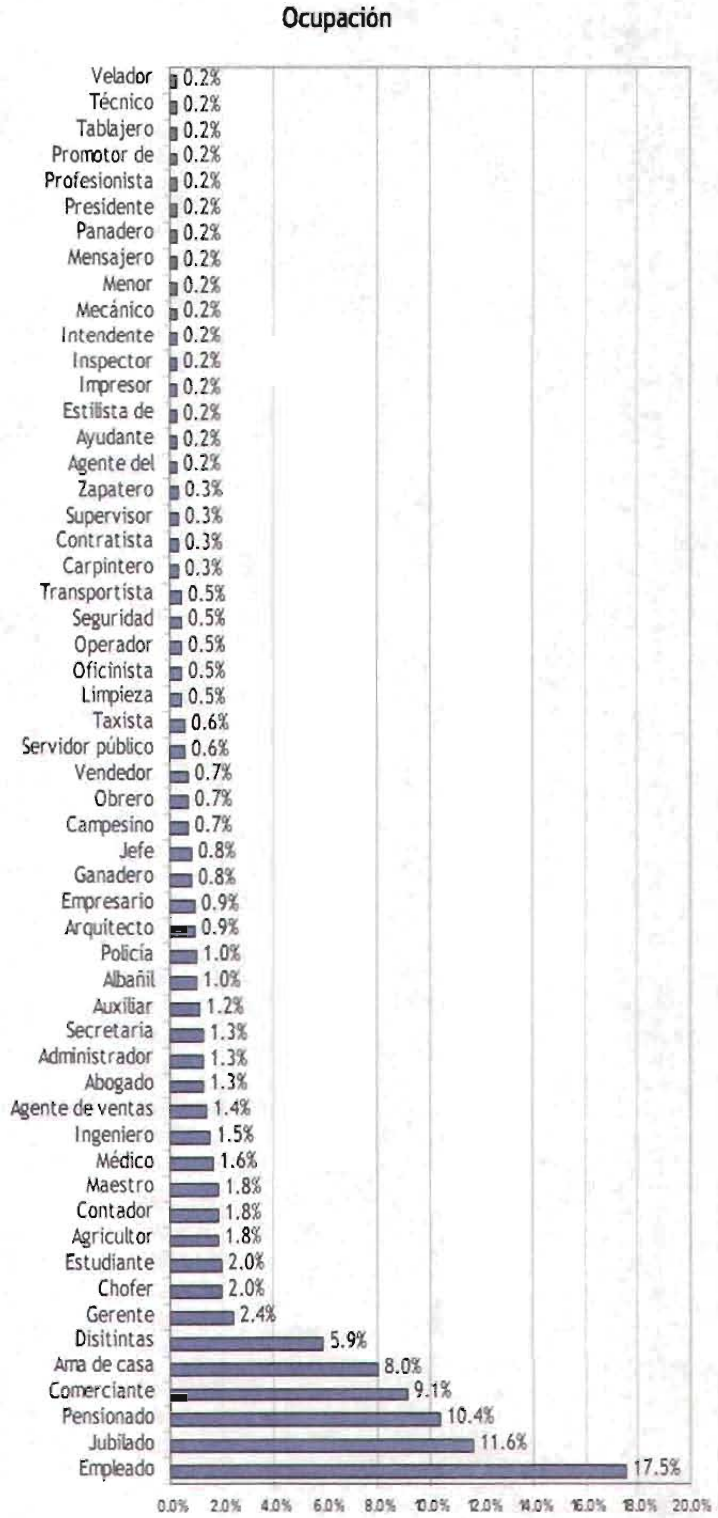
Distribución porcentual del número de siniestros de vida individual en la República Mexicana

Número de siniestros	
Estado	%
Distrito Federal	21.1%
Sin información	8.6%
Jalisco	7.1%
Chihuahua	5.6%
Sinaloa	5.4%
Veracruz	5.3%
Puebla	4.7%
Sonora	4.5%
Nuevo León	4.5%
Guanajuato	4.2%
Tamaulipas	3.4%
Coahuila	2.8%
Oaxaca	2.4%
Chiapas	2.3%
Michoacán	2.2%
Aguascalientes	2.1%
Durango	1.9%
Baja California	1.9%
Yucatán	1.4%
Tabasco	1.2%
Extranjero	1.0%
Morelos	0.9%
Querétaro	0.8%
Nayarit	0.8%
Zacatecas	0.7%
San Luis Potosí	0.7%
Guerrero	0.6%
Colima	0.6%
Quintana Roo	0.5%
Campeche	0.3%
Hidalgo	0.2%
Baja California Sur	0.2%
Tlaxcala	0.1%
Total general	100.0%

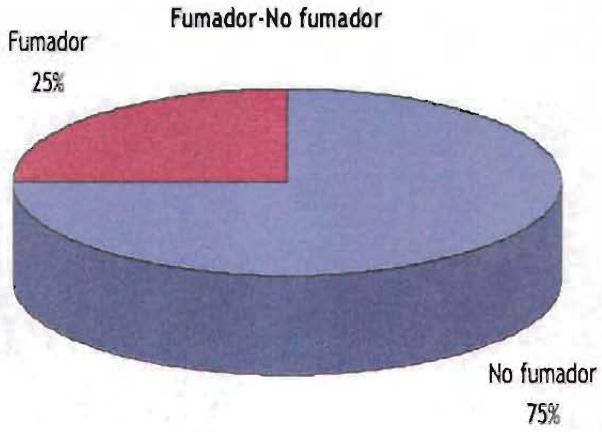


Distribución porcentual de las ocupaciones en la cartera de siniestros de vida individual

Ocupación	%
Agente del ministerio público	0.2%
Ayudante	0.2%
Estilista de belleza	0.2%
Impresor	0.2%
Inspector	0.2%
Intendente	0.2%
Mecánico	0.2%
Menor	0.2%
Mensajero	0.2%
Panadero	0.2%
Presidente municipal	0.2%
Profesionista	0.2%
Promotor de seguros	0.2%
Tablajero	0.2%
Técnico	0.2%
Velador	0.2%
Carpintero	0.3%
Contratista	0.3%
Supervisor	0.3%
Zapatero	0.3%
Limpieza	0.5%
Oficinista	0.5%
Operador	0.5%
Seguridad	0.5%
Transportista	0.5%
Servidor público	0.6%
Taxista	0.6%
Campesino	0.7%
Obrero	0.7%
Vendedor	0.7%
Ganadero	0.8%
Jefe	0.8%
Arquitecto	0.9%
Empresario	0.9%
Albañil	1.0%
Policía	1.0%
Auxiliar	1.2%
Abogado	1.3%
Administrador	1.3%
Secretaria	1.3%
Agente de ventas	1.4%
Ingeniero	1.5%
Médico	1.6%
Agricultor	1.8%
Contador	1.8%
Maestro	1.8%
Chofer	2.0%
Estudiante	2.0%
Gerente	2.4%
Disitintas ocupaciones con 1 siniestro	5.9%
Ama de casa	8.0%
Comerciante	9.1%
Pensionado	10.4%
Jubilado	11.6%
Empleado	17.5%
Total general	**100.0%**

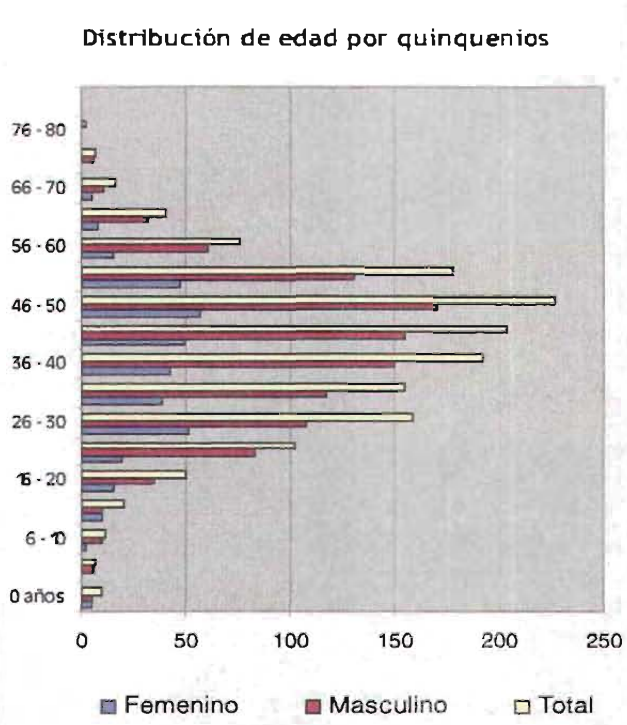


Distribución de la cartera por fumador/ no fumador



Los siniestros reportados durante el año, presentan la siguiente distribución por grupos de edad, en la cual puede observarse que la mayoría está por encima de la edad promedio de la cartera, que es 38 años.

Quinquenio	Femenino	Masculino	Total
0 años	5	5	10
1 - 5	1	6	7
6 - 10	2	10	12
11 - 15	10	10	20
16 - 20	15	35	50
21 - 25	19	83	102
26 - 30	51	108	159
31 - 35	38	117	155
36 - 40	42	150	192
41 - 45	49	155	204
46 - 50	57	170	227
51 - 55	47	131	178
56 - 60	15	61	76
61 - 65	8	32	40
66 - 70	5	11	16
71 - 75	1	6	7
76 - 80	1	1	2
81 - 85	0	1	1
Total	366	1092	1458



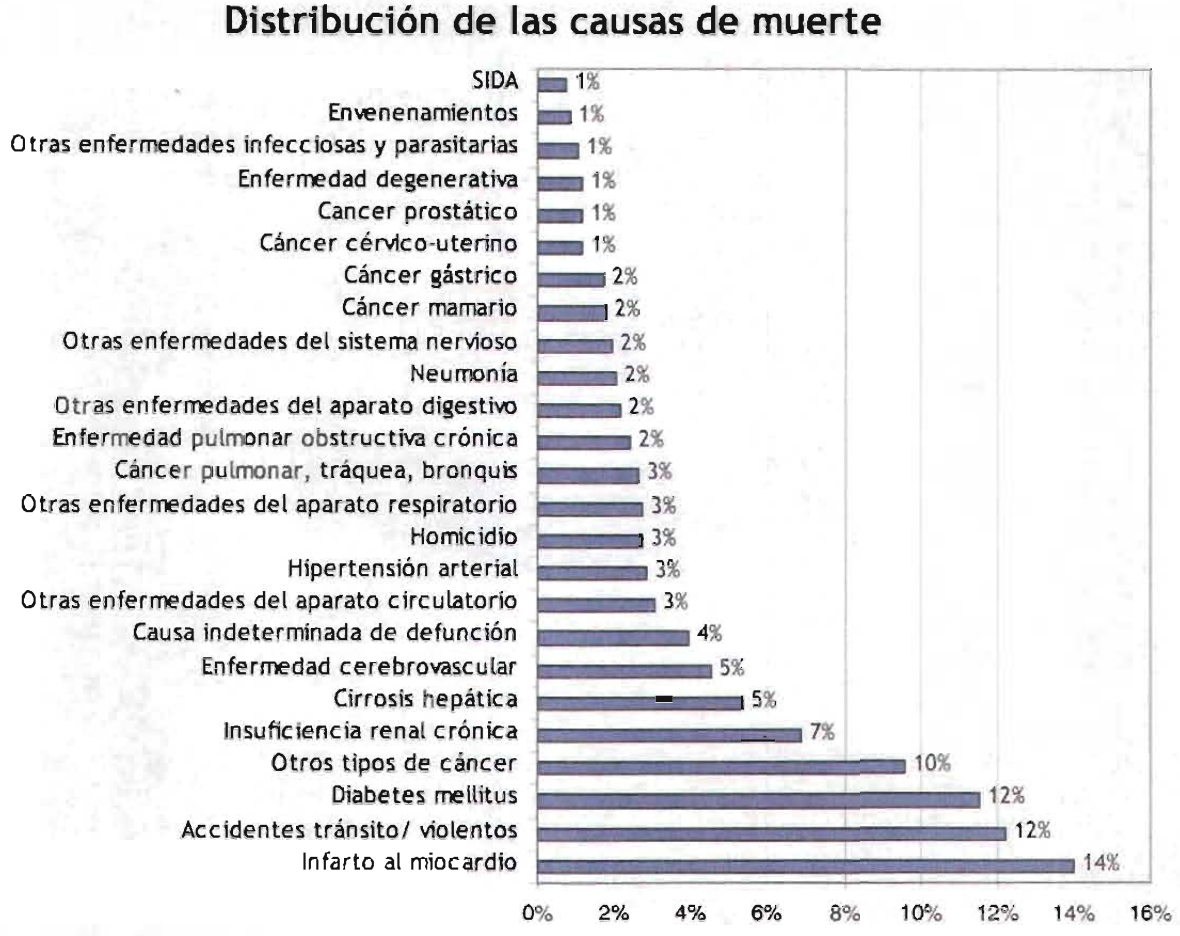
5.2 Estadísticas descriptivas

Para el presente ejercicio, se están considerando los siniestros anuales de una cartera de vida individual, divididos por sexo y considerando su edad promedio.

Esta cartera contempla 25 principales causas de muerte:

Causa	Numero de siniestros		
	Total	%	% Acum
Infarto al miocardio	204	14%	14%
Accidentes tránsito/ violentos	178	12%	26%
Diabetes mellitus	168	12%	38%
Otros tipos de cáncer	140	10%	47%
Insuficiencia renal crónica	100	7%	54%
Cirrosis hepática	78	5%	60%
Enfermedad cerebrovascular	66	5%	64%
Causa indeterminada de defunción	57	4%	68%
Otras enfermedades del aparato circulatorio	44	3%	71%
Hipertensión arterial	41	3%	74%
Homicidio	40	3%	77%
Otras enfermedades del aparato respiratorio	40	3%	79%
Cáncer pulmonar, tráquea, bronquis	38	3%	82%
Enfermedad pulmonar obstructiva crónica	35	2%	84%
Otras enfermedades del aparato digestivo	31	2%	86%
Neumonía	30	2%	88%
Otras enfermedades del sistema nervioso	28	2%	90%
Cáncer mamario	26	2%	92%
Cáncer gástrico	25	2%	94%
Cáncer cérvico-uterino	17	1%	95%
Cancer prostático	17	1%	96%
Enfermedad degenerativa	17	1%	97%
Otras enfermedades infecciosas y parasitarias	15	1%	98%
Envenenamientos	12	1%	99%
SIDA	11	1%	100%
Total general	***1,458***	***100%***	

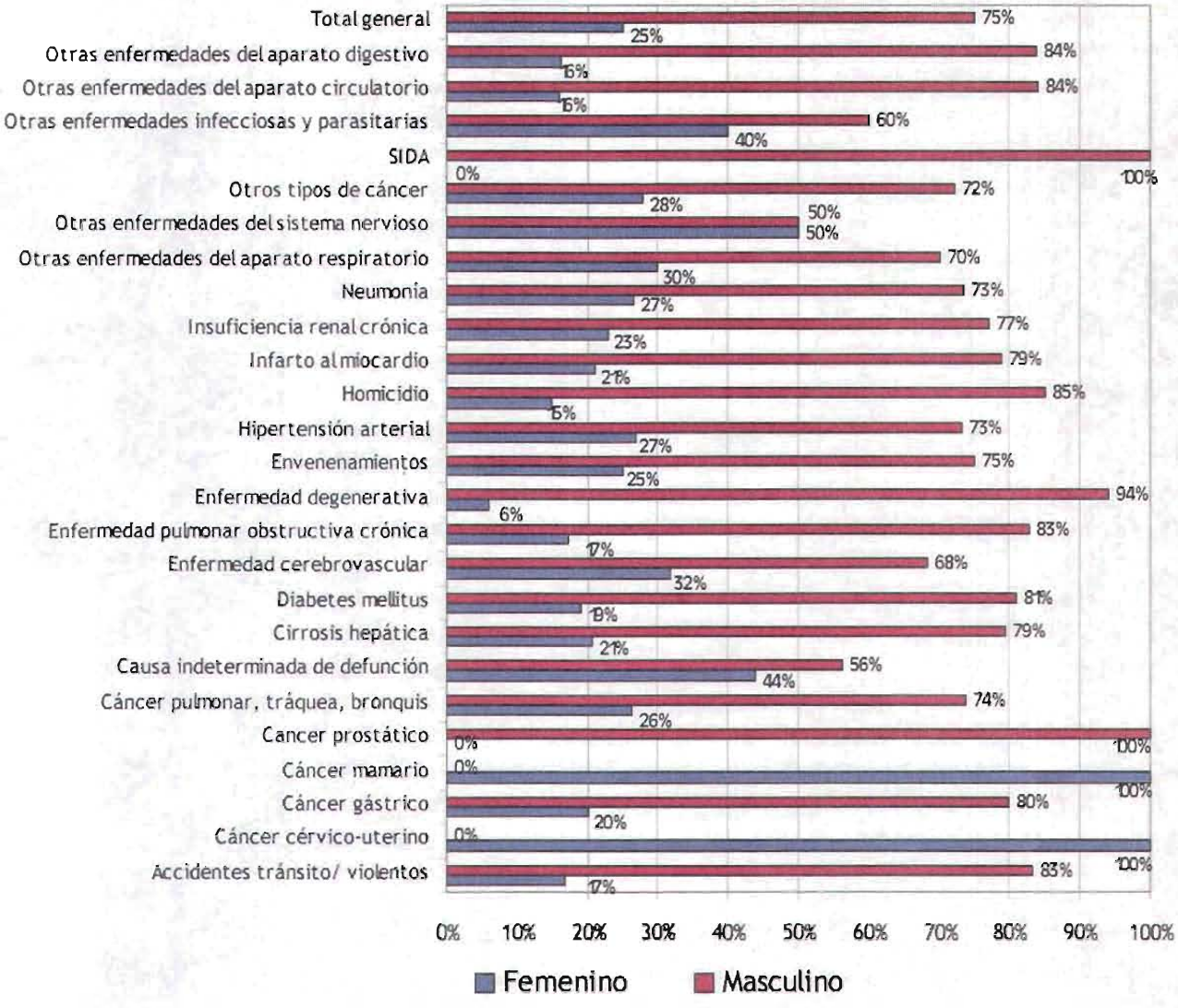
La distribución de estás causas puede observarse más fácilmente en la siguiente gráfica:



La causa de muerte más importante para toda la cartera es el infarto al miocardio, con 14%, mientras que el SIDA, los envenenamientos, las enfermedades infecciosas y las parasitarias, la enfermedad degenerativa, el cáncer prostático y el cáncer cérvico-uterino alcanzan el 1% cada uno.

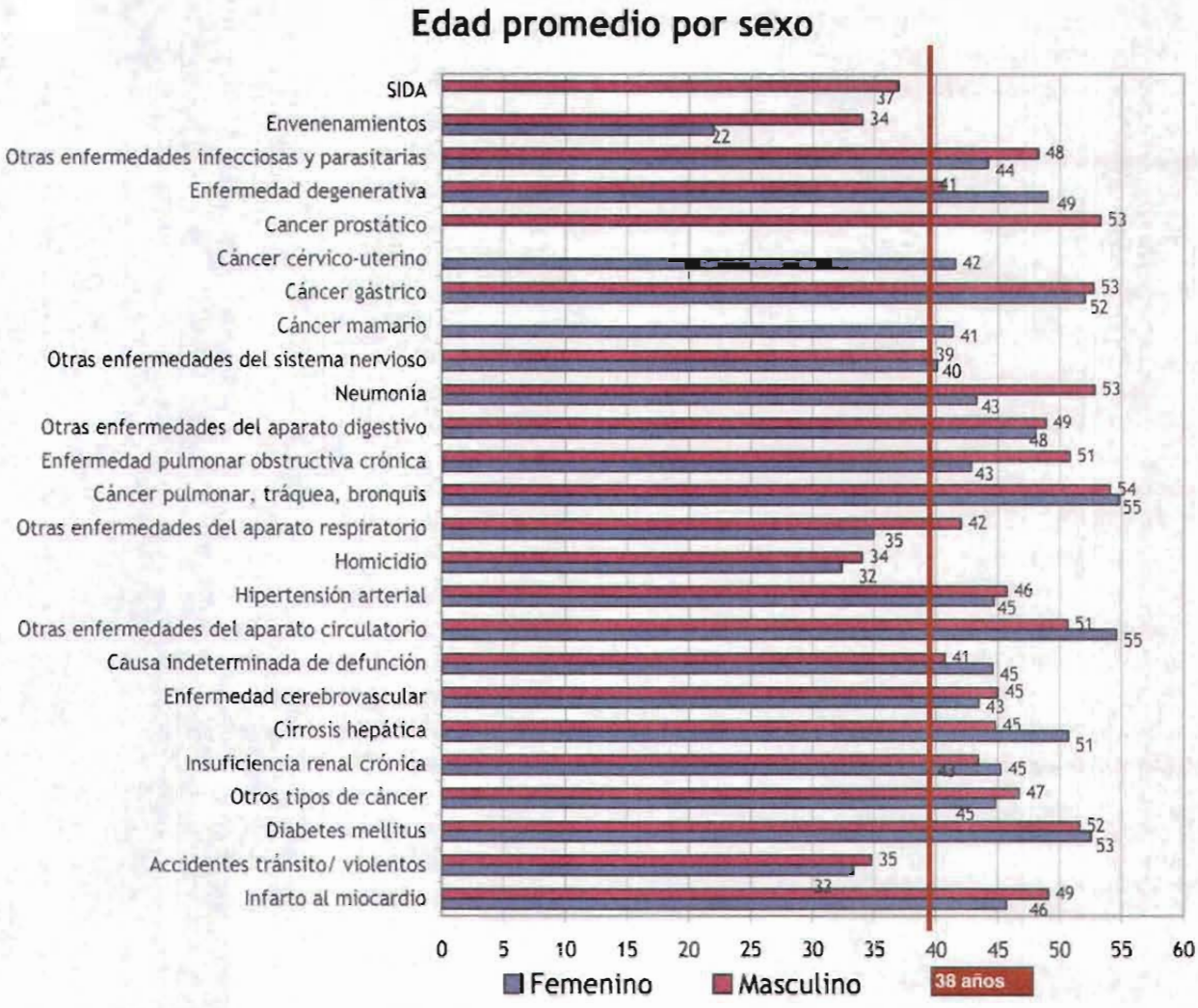
Puede revisarse también, de estos casos, la distribución de causas de muerte por sexo.

Distribución de las causas de muerte por sexo



Se observa que el 75% de los siniestros pertenecen al sexo masculino. Sin considerar las causas de muerte propias al sexo, el número de siniestros ocurridos al sexo masculino es mayor en todas las causas al número de siniestros ocurridos del sexo femenino.

Considerando la edad promedio a la fecha de ocurrencia del siniestro, tenemos la siguiente distribución:



La edad promedio de la cartera de vida individual es 38 años, las causas de muerte como SIDA (masculino), envenenamientos (masculino y femenino), otras enfermedades del aparato respiratorio (femenino), homicidios y accidentes (masculino y femenino), tienen la edad promedio por debajo de la correspondiente a la cartera.

5.3 Resultados del Modelo Log-Lineal de 2x2

El primer modelo que se revisará será la causa de muerte contra el sexo. Se comenzará revisando la independencia entre las dos variables, para determinar el modelo que se usará.

Se obtuvo una $\chi^2 = 51.0753$, con un nivel de significancia del 5%, $\chi_{.05}^2 = 36.4151$, como 36.4151< 51.0757, se rechaza la hipótesis de independencia.

Se utilizará el modelo saturado: $\ln E_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$

La estimación se basa en la tabla de contingencia: $E_{ij} = n_{ij}$

Causa	Número de siniestros		
	Femenino	Masculino	Total
Infarto al miocardio	43	161	204
Accidentes tránsito/ violentos	30	148	178
Diabetes mellitus	32	136	168
Otros tipos de cáncer	39	101	140
Insuficiencia renal crónica	23	77	100
Cirrosis hepática	16	62	78
Enfermedad cerebrovascular	21	45	66
Causa indeterminada de defunción	25	32	57
Otras enfermedades del aparato circulatorio	7	37	44
Hipertensión arterial	11	30	41
Homicidio	6	34	40
Otras enfermedades del aparato respiratorio	12	28	40
Cáncer pulmonar, tráquea, bronquis	10	28	38
Enfermedad pulmonar obstructiva crónica	6	29	35
Otras enfermedades del aparato digestivo	5	26	31
Neumonía	8	22	30
Otras enfermedades del sistema nervioso	14	14	28
Cáncer mamario	26	0	26
Cáncer gástrico	5	20	25
Cáncer cérvico-uterino	17	0	17
Cáncer prostático	0	17	17
Enfermedad degenerativa	1	16	17
Otras enfermedades infecciosas y parasitarias	6	9	15
Envenenamientos	3	9	12
SIDA	0	11	11
Total general	*366*	*1092*	*1458*

Aplicando el logaritmo a la tabla:

Causa	Número de siniestros		
	Femenino	Masculino	Total
Infarto al miocardio	3.7612	5.0814	8.8426
Accidentes tránsito/ violentos	3.4012	4.9972	8.3984
Diabetes mellitus	3.4657	4.9127	8.3784
Otros tipos de cáncer	3.6636	4.6151	8.2787
Insuficiencia renal crónica	3.1355	4.3438	7.4793
Cirrosis hepática	2.7726	4.1271	6.8997
Enfermedad cerebrovascular	3.0445	3.8067	6.8512
Causa indeterminada de defunción	3.2189	3.4657	6.6846
Otras enfermedades del aparato circulatorio	1.9459	3.6109	5.5568
Hipertensión arterial	2.3979	3.4012	5.7991
Homicidio	1.7918	3.5264	5.3181
Otras enfermedades del aparato respiratorio	2.4849	3.3322	5.8171
Cáncer pulmonar, tráquea, bronquis	2.3026	3.3322	5.6348
Enfermedad pulmonar obstructiva crónica	1.7918	3.3673	5.1591
Otras enfermedades del aparato digestivo	1.6094	3.2581	4.8675
Neumonía	2.0794	3.0910	5.1705
Otras enfermedades del sistema nervioso	2.6391	2.6391	5.2781
Cáncer mamario	3.2581	0.0000	3.2581
Cáncer gástrico	1.6094	2.9957	4.6052
Cáncer cérvico-uterino	2.8332	0.0000	2.8332
Cancer prostático	0.0000	2.8332	2.8332
Enfermedad degenerativa	0.0000	2.7726	2.7726
Otras enfermedades infecciosas y parasitarias	1.7918	2.1972	3.9890
Envenenamientos	1.0986	2.1972	3.2958
SIDA	0.0000	2.3979	2.3979
Total general	*56.0970*	*80.3020*	*136.3990*

El total es la suma de las causas por sexo, no se les aplica el logaritmo.

Los estimadores del modelo son:

$$\hat{u} = \frac{1}{rc}\sum n_{ij}$$

$$\hat{u}_{1(i)} = \frac{1}{c}\sum \ln n_{ij} - \hat{u}$$

$$\hat{u}_{2(j)} = \frac{1}{r}\sum \ln n_{ij} - \hat{u}$$

$r = 25$, $c = 2$ $\qquad$ $\hat{u} = 2.7280$

Estimadores

Renglones		Columnas	
$\hat{u}_{1(1)}$ =	1.6933	$\hat{u}_{2(1)}$ =	-0.4841
$\hat{u}_{1(2)}$ =	1.4712	$\hat{u}_{2(2)}$ =	0.4841
$\hat{u}_{1(3)}$ =	1.4612		
$\hat{u}_{1(4)}$ =	1.4114		
$\hat{u}_{1(5)}$ =	1.0117		
$\hat{u}_{1(6)}$ =	0.7219		
$\hat{u}_{1(7)}$ =	0.6976		
$\hat{u}_{1(8)}$ =	0.6143		
$\hat{u}_{1(9)}$ =	0.0504		
$\hat{u}_{1(10)}$ =	0.1716		
$\hat{u}_{1(11)}$ =	-0.0689		
$\hat{u}_{1(12)}$ =	0.1806		
$\hat{u}_{1(13)}$ =	0.0894		
$\hat{u}_{1(14)}$ =	-0.1485		
$\hat{u}_{1(15)}$ =	-0.2942		
$\hat{u}_{1(16)}$ =	-0.1427		
$\hat{u}_{1(17)}$ =	-0.0889		
$\hat{u}_{1(18)}$ =	-1.0989		
$\hat{u}_{1(19)}$ =	-0.4254		
$\hat{u}_{1(20)}$ =	-1.3114		
$\hat{u}_{1(21)}$ =	-1.3114		
$\hat{u}_{1(22)}$ =	-1.3417		
$\hat{u}_{1(23)}$ =	-0.7335		
$\hat{u}_{1(24)}$ =	-1.0801		
$\hat{u}_{1(25)}$ =	-1.5290		

Efectos

Renglones		Columnas	
$\partial_{1(1)}$ =	5.4375	$\partial_{2(1)}$ =	0.6163
$\partial_{1(2)}$ =	4.3546	$\partial_{2(2)}$ =	1.6227
$\partial_{1(3)}$ =	4.3112		
$\partial_{1(4)}$ =	4.1015		
$\partial_{1(5)}$ =	2.7502		
$\partial_{1(6)}$ =	2.0583		
$\partial_{1(7)}$ =	2.0089		
$\partial_{1(8)}$ =	1.8484		
$\partial_{1(9)}$ =	1.0517		
$\partial_{1(10)}$ =	1.1872		
$\partial_{1(11)}$ =	0.9334		
$\partial_{1(12)}$ =	1.1979		
$\partial_{1(13)}$ =	1.0935		
$\partial_{1(14)}$ =	0.8620		
$\partial_{1(15)}$ =	0.7451		
$\partial_{1(16)}$ =	0.8670		
$\partial_{1(17)}$ =	0.9149		
$\partial_{1(18)}$ =	0.3332		
$\partial_{1(19)}$ =	0.6535		
$\partial_{1(20)}$ =	0.2694		
$\partial_{1(21)}$ =	0.2694		
$\partial_{1(22)}$ =	0.2614		
$\partial_{1(23)}$ =	0.4802		
$\partial_{1(24)}$ =	0.3396		
$\partial_{1(25)}$ =	0.2167		

Los efectos son las exponenciales de los estimadores.

De los efectos sobre el número de siniestros por causa de muerte, el correspondiente al marcado $\partial_{1(25)} = .2167$, SIDA, es el que tiene una influencia menor, mientras que $\partial_{1(1)} = 5.4375$, Infarto al miocardio, es el más alto.

Respecto al efecto del sexo, el que tiene menor influencia es el sexo femenino, , $\partial_{2(1)} = .6163$ mientras que el masculino tiene mayor influencia, $\partial_{2(2)} = 1.6227$.

Obteniendo las iteracciones de la tabla:

Causa	Número de siniestros		
	Femenino	Masculino	Total
Infarto al miocardio	-0.1760	0.1760	0.0000
Accidentes tránsito/ violentos	-0.3139	0.3139	0.0000
Diabetes mellitus	-0.2394	0.2394	0.0000
Otros tipos de cáncer	0.0083	-0.0083	0.0000
Insuficiencia renal crónica	-0.1201	0.1201	0.0000
Cirrosis hepática	-0.1932	0.1932	0.0000
Enfermedad cerebrovascular	0.1030	-0.1030	0.0000
Causa indeterminada de defunción	0.3607	-0.3607	0.0000
Otras enfermedades del aparato circulatorio	-0.3484	0.3484	0.0000
Hipertensión arterial	-0.0176	0.0176	0.0000
Homicidio	-0.3832	0.3832	0.0000
Otras enfermedades del aparato respiratorio	0.0604	-0.0604	0.0000
Cáncer pulmonar, tráquea, bronquis	-0.0307	0.0307	0.0000
Enfermedad pulmonar obstructiva crónica	-0.3037	0.3037	0.0000
Otras enfermedades del aparato digestivo	-0.3402	0.3402	0.0000
Neumonía	-0.0217	0.0217	0.0000
Otras enfermedades del sistema nervioso	0.4841	-0.4841	0.0000
Cáncer mamario	2.1131	-2.1131	0.0000
Cáncer gástrico	-0.2090	0.2090	0.0000
Cáncer cérvico-uterino	1.9007	-1.9007	0.0000
Cancer prostático	-0.9325	0.9325	0.0000
Enfermedad degenerativa	-0.9022	0.9022	0.0000
Otras enfermedades infecciosas y parasitarias	0.2814	-0.2814	0.0000
Envenenamientos	-0.0652	0.0652	0.0000
SIDA	-0.7148	0.7148	0.0000
Total general	*0.0000*	*0.0000*	*0.0000*

Calculando los efectos de la tabla, aplicando exponencial:

Causa	Número de siniestros	
	Femenino	Masculino
Infarto al miocardio	0.8386	1.1924
Accidentes tránsito/ violentos	0.7306	1.3688
Diabetes mellitus	0.7871	1.2704
Otros tipos de cáncer	1.0084	0.9917
Insuficiencia renal crónica	0.8869	1.1276
Cirrosis hepática	0.8243	1.2131
Enfermedad cerebrovascular	1.1085	0.9021
Causa indeterminada de defunción	1.4343	0.6972
Otras enfermedades del aparato circulatorio	0.7058	1.4168
Hipertensión arterial	0.9826	1.0177
Homicidio	0.6817	1.4670
Otras enfermedades del aparato respiratorio	1.0623	0.9413
Cáncer pulmonar, tráquea, bronquis	0.9698	1.0312
Enfermedad pulmonar obstructiva crónica	0.7381	1.3548
Otras enfermedades del aparato digestivo	0.7116	1.4053
Neumonía	0.9785	1.0219
Otras enfermedades del sistema nervioso	1.6227	0.6163
Cáncer mamario	8.2742	0.1209
Cáncer gástrico	0.8114	1.2325
Cáncer cérvico-uterino	6.6906	0.1495
Cancer prostático	0.3936	2.5409
Enfermedad degenerativa	0.4057	2.4650
Otras enfermedades infecciosas y parasitarias	1.3249	0.7548
Envenenamientos	0.9369	1.0674
SIDA	0.4893	2.0439

En esta tabla, se obtiene la combinación de variables, causa de muerte y sexo, la relación se muestra sobre el número de siniestros, se observa que del sexo femenino, las causas que manifiestan un incremento en la frecuencia en un factor, son cáncer mamario y cáncer cérvico-uterino, por encima de la media general $\hat{u} = 2.2780$. Del sexo masculino, es el cáncer prostático y enfermedad degenerativa.

5.4 Resultados del Modelo Log-Lineal de 3x3

Variables:

i = *Sexo* $\quad i_1 = Femenino \quad i_2 = Masculino$

j= Edad promedio por debajo de la edad promedio de la cartera

$j_1 = Si \quad j_2 = No$

k= Causa de muerte

Causa de muerte	i_1 = Femenino			i_2 = Masculino			Total		
	j_1 = No	j_2 = Si	Total	j_1 = No	j_2 = Si	Total	No	Si	Total general
k_1 = Infarto al miocardio	32	11	43 n_{11}	142	19	161 n_{21}	174 $n_{\cdot 1}$	30 $n_{\cdot 2}$	204 $n_{\cdot\cdot}$
k_2 = Accidentes tránsito/ violentos	10	20	30 n_{12}	63	85	148 n_{22}	73 $n_{\cdot 1}$	105 $n_{\cdot 2}$	178 $n_{\cdot\cdot}$
k_3 = Diabetes mellitus	28	4	32 n_{13}	128	8	136 n_{23}	156 $n_{\cdot 1}$	12 $n_{\cdot 2}$	168 $n_{\cdot\cdot}$
k_4 = Otros tipos de cáncer	28	11	39 n_{14}	78	23	101 n_{24}	106 $n_{\cdot 1}$	34 $n_{\cdot 2}$	140 $n_{\cdot\cdot}$
k_5 = Insuficiencia renal crónica	16	7	23 n_{15}	55	22	77 n_{25}	71 $n_{\cdot 1}$	29 $n_{\cdot 2}$	100 $n_{\cdot\cdot}$
k_6 = Cirrosis hepática	15	1	16 n_{16}	49	13	62 n_{26}	64 $n_{\cdot 1}$	14 $n_{\cdot 2}$	78 $n_{\cdot\cdot}$
k_7 = Enfermedad cerebrovascular	16	5	21 n_{17}	32	13	45 n_{27}	48 $n_{\cdot 1}$	18 $n_{\cdot 2}$	66 $n_{\cdot\cdot}$
k_8 = Causa indeterminada de defunción	16	9	25 n_{18}	20	12	32 n_{28}	36 $n_{\cdot 1}$	21 $n_{\cdot 2}$	57 $n_{\cdot\cdot}$
k_9 = Otras enfermedades del aparato circulatorio	6	1	7 n_{19}	34	3	37 n_{29}	40 $n_{\cdot 1}$	4 $n_{\cdot 2}$	44 $n_{\cdot\cdot}$
k_{10} = Hipertensión arterial	8	3	11 n_{110}	22	8	30 n_{210}	30 $n_{\cdot 10}$	11 $n_{\cdot 20}$	41 $n_{\cdot\cdot}$
k_{11} = Homicidio	2	4	6 n_{111}	13	21	34 n_{211}	15 $n_{\cdot 11}$	25 $n_{\cdot 21}$	40 $n_{\cdot\cdot}$
k_{12} = Otras enfermedades del aparato respiratorio	6	6	12 n_{112}	20	8	28 n_{212}	26 $n_{\cdot 12}$	14 $n_{\cdot 22}$	40 $n_{\cdot\cdot}$
k_{13} = Cáncer pulmonar, tráquea, bronquis	10	0	10 n_{113}	26	2	28 n_{213}	36 $n_{\cdot 13}$	2 $n_{\cdot 23}$	38 $n_{\cdot\cdot}$
k_{14} = Enfermedad pulmonar obstructiva crónica	5	1	6 n_{114}	26	3	29 n_{214}	31 $n_{\cdot 14}$	4 $n_{\cdot 24}$	35 $n_{\cdot\cdot}$
k_{15} = Otras enfermedades del aparato digestivo	3	2	5 n_{115}	21	5	26 n_{215}	24 $n_{\cdot 15}$	7 $n_{\cdot 25}$	31 $n_{\cdot\cdot}$
k_{16} = Neumonía	6	2	8 n_{116}	20	2	22 n_{216}	26 $n_{\cdot 16}$	4 $n_{\cdot 26}$	30 $n_{\cdot\cdot}$
k_{17} = Otras enfermedades del sistema nervioso	6	8	14 n_{117}	7	7	14 n_{217}	13 $n_{\cdot 17}$	15 $n_{\cdot 27}$	28 $n_{\cdot\cdot}$
k_{18} = Cáncer mamario	13	13	26 n_{118}	0	0	0 n_{218}	13 $n_{\cdot 18}$	13 $n_{\cdot 28}$	26 $n_{\cdot\cdot}$
k_{19} = Cáncer gástrico	5	0	5 n_{119}	19	1	20 n_{219}	24 $n_{\cdot 19}$	1 $n_{\cdot 29}$	25 $n_{\cdot\cdot}$
k_{20} = Cáncer cérvico-uterino	13	4	17 n_{120}	0	0	0 n_{220}	13 $n_{\cdot 20}$	4 $n_{\cdot 30}$	17 $n_{\cdot\cdot}$
k_{21} = Cancer prostático	0	0	0 n_{121}	17		17 n_{221}	17 $n_{\cdot 21}$	0 $n_{\cdot 31}$	17 $n_{\cdot\cdot}$
k_{22} = Enfermedad degenerativa	1	0	1 n_{122}	8	8	16 n_{222}	9 $n_{\cdot 22}$	8 $n_{\cdot 32}$	17 $n_{\cdot\cdot}$
k_{23} = Otras enfermedades infecciosas y parasitarias	4	2	6 n_{123}	7	2	9 n_{223}	11 $n_{\cdot 23}$	4 $n_{\cdot 33}$	15 $n_{\cdot\cdot}$
k_{24} = Envenenamientos	0	3	3 n_{124}		9	9 n_{224}	0 $n_{\cdot 24}$	12 $n_{\cdot 34}$	12 $n_{\cdot\cdot}$
k_{25} = SIDA	0	0	0 n_{125}	5	6	11 n_{225}	5 $n_{\cdot 25}$	6 $n_{\cdot 35}$	11 $n_{\cdot\cdot}$
Total general	*249* n_{11}	*117* n_{12}	*366* n_1	*812* n_{21}	*280* n_{22}	*1,092* n_2	*1,061* n_1	*397* n_2	*1,458* n

De la tabla anterior, se aplica el logaritmo obteniendo los siguientes resultados:

Totales:

$n_{1..}$	=	366	ln(366)	=	5.9026
$n_{2..}$	=	1,092	ln(1092)	=	6.9958
$n_{.1.}$	=	1,061	ln(1061)	=	6.9670
$n_{.2.}$	=	397	ln(397)	=	5.9839
$n_{..1}$	=	204	ln(204)	=	5.3181
$n_{..2}$	=	178	ln(178)	=	5.1818
$n_{..3}$	=	168	ln(168)	=	5.1240
$n_{..4}$	=	140	ln(140)	=	4.9416
$n_{..5}$	=	100	ln(100)	=	4.6052
$n_{..6}$	=	78	ln(78)	=	4.3567
$n_{..7}$	=	66	ln(66)	=	4.1897
$n_{..8}$	=	57	ln(57)	=	4.0431
$n_{..9}$	=	44	ln(44)	=	3.7842
$n_{..10}$	=	41	ln(41)	=	3.7136
$n_{..11}$	=	40	ln(40)	=	3.6889
$n_{..12}$	=	40	ln(40)	=	3.6889
$n_{..13}$	=	38	ln(38)	=	3.6376
$n_{..14}$	=	35	ln(35)	=	3.5553
$n_{..15}$	=	31	ln(31)	=	3.4340
$n_{..16}$	=	30	ln(30)	=	3.4012
$n_{..17}$	=	28	ln(28)	=	3.3322
$n_{..18}$	=	26	ln(26)	=	3.2581
$n_{..19}$	=	25	ln(25)	=	3.2189
$n_{..20}$	=	17	ln(17)	=	2.8332
$n_{..21}$	=	17	ln(17)	=	2.8332
$n_{..22}$	=	17	ln(17)	=	2.8332
$n_{..23}$	=	15	ln(15)	=	2.7081
$n_{..24}$	=	12	ln(12)	=	2.4849
$n_{..25}$	=	11	ln(11)	=	2.3979

Subtotales para la variable *j*:

$n_{11.}$	=	249	ln(249)	=	5.5175	$n_{21.}$	=	812	ln(812)	=	6.6995
n_{12}	=	117	ln(117)	=	4.7622	$n_{22.}$	=	280	ln(280)	=	5.6348
$n_{.11}$	=	174	ln(174)	=	5.1591	$n_{.21}$	=	30	ln(30)	=	3.4012
$n_{.12}$	=	73	ln(73)	=	4.2905	$n_{.22}$	=	105	ln(105)	=	4.6540
$n_{.13}$	=	156	ln(156)	=	5.0499	$n_{.23}$	=	12	ln(12)	=	2.4849
$n_{.14}$	=	106	ln(106)	=	4.6634	$n_{.24}$	=	34	ln(34)	=	3.5264
$n_{.15}$	=	71	ln(71)	=	4.2627	$n_{.25}$	=	29	ln(29)	=	3.3673
$n_{.16}$	=	64	ln(64)	=	4.1589	$n_{.26}$	=	14	ln(14)	=	2.6391
$n_{.17}$	=	48	ln(48)	=	3.8712	$n_{.27}$	=	18	ln(18)	=	2.8904
$n_{.18}$	=	36	ln(36)	=	3.5835	$n_{.28}$	=	21	ln(21)	=	3.0445
$n_{.19}$	=	40	ln(40)	=	3.6889	$n_{.29}$	=	4	ln(4)	=	1.3863
$n_{.110}$	=	30	ln(30)	=	3.4012	$n_{.210}$	=	11	ln(11)	=	2.3979
$n_{.111}$	=	15	ln(15)	=	2.7081	$n_{.211}$	=	25	ln(25)	=	3.2189
$n_{.112}$	=	26	ln(26)	=	3.2581	$n_{.212}$	=	14	ln(14)	=	2.6391
$n_{.113}$	=	36	ln(36)	=	3.5835	$n_{.213}$	=	2	ln(2)	=	0.6931
$n_{.114}$	=	31	ln(31)	=	3.4340	$n_{.214}$	=	4	ln(4)	=	1.3863
$n_{.115}$	=	24	ln(24)	=	3.1781	$n_{.215}$	=	7	ln(7)	=	1.9459
$n_{.116}$	=	26	ln(26)	=	3.2581	$n_{.216}$	=	4	ln(4)	=	1.3863
$n_{.117}$	=	13	ln(13)	=	2.5649	$n_{.217}$	=	15	ln(15)	=	2.7081
$n_{.118}$	=	13	ln(13)	=	2.5649	$n_{.218}$	=	13	ln(13)	=	2.5649
$n_{.119}$	=	24	ln(24)	=	3.1781	$n_{.219}$	=	1	ln(1)	=	0.0000
$n_{.120}$	=	13	ln(13)	=	2.5649	$n_{.220}$	=	4	ln(4)	=	1.3863
$n_{.121}$	=	17	ln(17)	=	2.8332	$n_{.221}$	=	0	ln(0)	=	?
$n_{.122}$	=	9	ln(9)	=	2.1972	$n_{.222}$	=	8	ln(8)	=	2.0794
$n_{.123}$	=	11	ln(11)	=	2.3979	$n_{.223}$	=	4	ln(4)	=	1.3863
$n_{.124}$	=	0	ln(0)	=	?	$n_{.224}$	=	12	ln(12)	=	2.4849
$n_{.125}$	=	5	ln(5)	=	1.6094	$n_{.225}$	=	6	ln(6)	=	1.7918

Subtotales para la variable *i*:

n_{11}	= 43	ln(43)	= 3.7612	n_{21}	= 161	ln(161)	= 5.08140
n_{12}	= 30	ln(30)	= 3.4012	n_{22}	= 148	ln(148)	= 4.99721
n_{13}	= 32	ln(32)	= 3.4657	n_{23}	= 136	ln(136)	= 4.91265
n_{14}	= 39	ln(39)	= 3.6636	n_{24}	= 101	ln(101)	= 4.61512
n_{15}	= 23	ln(23)	= 3.1355	n_{25}	= 77	ln(77)	= 4.34381
n_{16}	= 16	ln(16)	= 2.7726	n_{26}	= 62	ln(62)	= 4.12713
n_{17}	= 21	ln(21)	= 3.0445	n_{27}	= 45	ln(45)	= 3.80666
n_{18}	= 25	ln(25)	= 3.2189	n_{28}	= 32	ln(32)	= 3.46574
n_{19}	= 7	ln(7)	= 1.9459	n_{29}	= 37	ln(37)	= 3.61092
n_{110}	= 11	ln(11)	= 2.3979	n_{210}	= 30	ln(30)	= 3.40120
n_{111}	= 6	ln(6)	= 1.7918	n_{211}	= 34	ln(34)	= 3.52636
n_{112}	= 12	ln(12)	= 2.4849	n_{212}	= 28	ln(28)	= 3.33220
n_{113}	= 10	ln(10)	= 2.3026	n_{213}	= 28	ln(28)	= 3.33220
n_{114}	= 6	ln(6)	= 1.7918	n_{214}	= 29	ln(29)	= 3.36730
n_{115}	= 5	ln(5)	= 1.6094	n_{215}	= 26	ln(26)	= 3.25810
n_{116}	= 8	ln(8)	= 2.0794	n_{216}	= 22	ln(22)	= 3.09104
n_{117}	= 14	ln(14)	= 2.6391	n_{217}	= 14	ln(14)	= 2.63906
n_{118}	= 26	ln(26)	= 3.2581	n_{218}	= 0	ln(0)	= ?
n_{119}	= 5	ln(5)	= 1.6094	n_{219}	= 20	ln(20)	= 2.99573
n_{120}	= 17	ln(17)	= 2.8332	n_{220}	= 0	ln(0)	= ?
n_{121}	= 0	ln(0)	= ?	n_{221}	= 17	ln(17)	= 2.83321
n_{122}	= 1	ln(1)	= 0.0000	n_{222}	= 16	ln(16)	= 2.77259
n_{123}	= 6	ln(6)	= 1.7918	n_{223}	= 9	ln(9)	= 2.19722
n_{124}	= 3	ln(3)	= 1.0986	n_{224}	= 9	ln(9)	= 2.19722
n_{125}	= 0	ln(0)	= ?	n_{225}	= 11	ln(11)	= 2.39790

Modelo de independencia parcial

Hipótesis:

$u_{13(ik)} = 0$ $\quad$ $u_{23(jk)} = 0$ $\quad$ $u_{123(ijk)} = 0$

Modelo: $LnE_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$

$$\hat{E}_{ijk} = \frac{n_{ij} n_{.k}}{N}$$

$\chi^2 = 489.25$, con 49 grados de libertad.

$\therefore$ Se rechaza la hipótesis de independencia.

$$\hat{u} = \frac{1}{rc}\sum_{i}^{r}\sum_{j}^{c} Ln(n_{ij}) + \frac{1}{s}\sum_{k=1}^{s} Ln(n_{..k}) - LnN$$

$$\hat{u}_{1(i)} = \frac{1}{c}\sum_{j=1}^{c} Ln(n_{ij}) - \frac{1}{rc}\sum_{i}^{r}\sum_{j}^{c} Ln(n_{ij})$$

$$\hat{u}_{2(j)} = \frac{1}{r}\sum_{i=1}^{r} Ln(n_{ij}) - \frac{1}{rc}\sum_{i}^{r}\sum_{j}^{c} Ln(n_{ij})$$

$$\hat{u}_{3(k)} = Ln(n_{..k}) - \frac{1}{s}\sum_{k=1}^{s} Ln(n_{..k})$$

$$\hat{u}_{12(ij)} = Ln(n_{ij}) + \frac{1}{s}\sum_{k=1}^{s} Ln(n_{..k}) - InN - \hat{u}_{1(i)} - \hat{u}_{2(j)} - \hat{u}$$

$$\hat{u} = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2} Ln(n_{ij}) + \frac{1}{8}\sum_{k=1}^{8} ln(n_{..k}) - Ln(n)$$

$$\hat{u} = \frac{1}{4}(\ln n_{11} + \ln n_{12.} + \ln n_{21.} + \ln n_{22.}) + \frac{1}{8}(\ln n_{..1} + \ln n_{..2} + \ln n_{..3} + \ln n_{..4} + \ln n_{..5} + \ln n_{..6} + \ln n_{..7} + \ln n_{..8} + \ln n_{..9} + \ln n_{..10} + \ln n_{..11} + \ln n_{..12} + \ln n_{..13} + \ln n_{..14} + \ln n_{..15} + \ln n_{..16} + \ln n_{..17} + \ln n_{..18} + \ln n_{..19} + \ln n_{..20} + \ln n_{..21} + \ln n_{..22} + \ln n_{..23} + \ln n_{..24} + \ln n_{..25}) - \ln n_{...}$$

$$\hat{u} = \frac{1}{4}(\ln 249 + \ln 117 + \ln 812 + \ln 280) + \frac{1}{8}(\ln 204 + \ln 178 + \ln 168 + \ln 140 + \ln 100 + \ln 78 + \ln 66 + \ln 57 + \ln 44 + \ln 41 + \ln 40 + \ln 40 + \ln 38 + \ln 35 + \ln 31 + \ln 30 + \ln 28 + \ln 26 + \ln 25 + \ln 17 + \ln 17 + \ln 17 + \ln 15 + \ln 12 + \ln 11) - \ln 1458$$

$$\hat{u} = \frac{1}{4}(22.6139) + \frac{1}{8}(92.5634) - 7.2848$$

$$\hat{u} = 9.9391$$

Aplicando el modelo a la variable $i = Sexo$

$$\hat{u}_{1(1)} = \frac{1}{c}(\ln n_{11} + \ln n_{12.}) - \frac{1}{rc}(\ln n_{11} + \ln n_{12} + \ln n_{21.} + \ln n_{22.})$$

$$\hat{u}_{1(1)} = \frac{1}{2}(\ln 249 + \ln 117) - \frac{1}{4}(\ln 249 + \ln 117 + \ln 812 + \ln 280)$$

$$\hat{u}_{1(1)} = \frac{1}{2}(5.5175 + 4.7622) - \frac{1}{4}(5.5175 + 4.7622 + 6.6995 + 5.6348)$$

$$\hat{u}_{1(1)} = -0.5137 \qquad \Rightarrow \partial_{1(1)} = e^{-0.5137} = 0.5983$$

$$\hat{u}_{2(1)} = \frac{1}{2}(6.6995 + 5.6348) - \frac{1}{4}(5.5175 + 4.7622 + 6.6975 + 5.6348)$$

$$\hat{u}_{1(2)} = 0.5137 \qquad \Rightarrow \partial_{1(1)} = e^{0.5137} = 1.6714$$

Se obtiene que el que tiene mayor influencia es el sexo masculino, con $\partial_{1(1)} = 1.6714$.

Aplicando el modelo a la *variable j= Edad promedio por debajo de la correspondiente a la de la cartera.*

$$j_1 = Si \qquad j_2 = No$$

$$\hat{u}_{2(1)} = \frac{1}{r}\sum_{i=1}^{r} \ln n_{ij} - \frac{1}{rc}\sum_{i}^{r}\sum_{j}^{c} \ln n_{ij}$$

$$\hat{u}_{2(1)} = \frac{1}{r}(\ln n_{11} + \ln n_{21}) - \frac{1}{rc}(\ln n_{11} + \ln n_{12} + \ln n_{21} + \ln n_{22})$$

$$\hat{u}_{2(1)} = \frac{1}{2}(\ln 249 + \ln 812) - \frac{1}{4}(\ln 249 + \ln 117 + \ln 812 + \ln 280)$$

$$\hat{u}_{2(1)} = \frac{1}{2}(5.5175 + 6.6995) - \frac{1}{4}(5.5175 + 4.7622 + 6.6995 + 5.6348)$$

$$\hat{u}_{2(1)} = 0.4550 \Rightarrow \partial_{2(1)} = e^{0.4550} = 1.5762$$

$$\hat{u}_{2(1)} = \frac{1}{2}(4.7622 + 5.6348) - \frac{1}{4}(5.5175 + 4.7622 + 6.6975 + 5.6348)$$

$$\hat{u}_{2(2)} = -0.4550 \qquad \Rightarrow \partial_{2(2)} = e^{-0.4550} = 0.6344$$

La más influyente corresponde a no estar por debajo de la edad promedio de la cartera con $\partial_{2(1)} = 1.5762$.

Respecto a la variable *k= Causa de muerte:*

$$\hat{u}_{3(k)} = \ln n_{..k} - \frac{1}{S}\sum_{k=1}^{s} \ln n_{..k}$$

$$\hat{u}_{3(1)} = \ln n_{..1} - \frac{1}{25}(\ln n_{..1} + \ln n_{..2} + \ln n_{..3} + \ln n_{..4} + \ln n_{..5} + \ln n_{..6} + \ln n_{..7} + \ln n_{..8} + \ln n_{..9} + \ln n_{..10} + \ln n_{..11} + \ln n_{..12} + \ln n_{..13} + \ln n_{..14} + \ln n_{..15} + \ln n_{..16} + \ln n_{..17} + \ln n_{..18} + \ln n_{..19} + \ln n_{..20} + \ln n_{..21} + \ln n_{..22} + \ln n_{..23} + \ln n_{..24} + \ln n_{..25})$$

$\hat{u}_{3(1)}$ =	5.3181 - $\frac{1}{25}$ (92.5634)=	1.6156	$\Rightarrow \partial_{3(1)} = e^{1.6156}$ =	5.0308
$\hat{u}_{3(2)}$ =	5.1818 - $\frac{1}{25}$ (92.5634)=	1.4792	$\Rightarrow \partial_{3(2)} = e^{1.4792}$ =	4.3896
$\hat{u}_{3(3)}$ =	5.1240 - $\frac{1}{25}$ (92.5634)=	1.4214	$\Rightarrow \partial_{3(3)} = e^{1.4214}$ =	4.1430
$\hat{u}_{3(4)}$ =	4.9416 - $\frac{1}{25}$ (92.5634)=	1.2391	$\Rightarrow \partial_{3(4)} = e^{1.2391}$ =	3.4525
$\hat{u}_{3(5)}$ =	4.6052 - $\frac{1}{25}$ (92.5634)=	0.9026	$\Rightarrow \partial_{3(5)} = e^{0.9026}$ =	2.4661
$\hat{u}_{3(6)}$ =	4.3567 - $\frac{1}{25}$ (92.5634)=	0.6542	$\Rightarrow \partial_{3(6)} = e^{0.6542}$ =	1.9236
$\hat{u}_{3(7)}$ =	4.1897 - $\frac{1}{25}$ (92.5634)=	0.4871	$\Rightarrow \partial_{3(7)} = e^{0.4871}$ =	1.6276
$\hat{u}_{3(8)}$ =	4.0431 - $\frac{1}{25}$ (92.5634)=	0.3405	$\Rightarrow \partial_{3(8)} = e^{0.3405}$ =	1.4057
$\hat{u}_{3(9)}$ =	3.7842 - $\frac{1}{25}$ (92.5634)=	0.0817	$\Rightarrow \partial_{3(9)} = e^{0.0817}$ =	1.0851
$\hat{u}_{3(10)}$ =	3.7136 - $\frac{1}{25}$ (92.5634)=	0.0110	$\Rightarrow \partial_{3(10)} = e^{0.0110}$ =	1.0111
$\hat{u}_{3(11)}$ =	3.6889 - $\frac{1}{25}$ (92.5634)=	-0.0137	$\Rightarrow \partial_{3(11)} = e^{-0.0137}$ =	0.9864
$\hat{u}_{3(12)}$ =	3.6889 - $\frac{1}{25}$ (92.5634)=	-0.0137	$\Rightarrow \partial_{3(12)} = e^{-0.0137}$ =	0.9864
$\hat{u}_{3(13)}$ =	3.6376 - $\frac{1}{25}$ (92.5634)=	-0.0649	$\Rightarrow \partial_{3(13)} = e^{-0.0649}$ =	0.9371
$\hat{u}_{3(14)}$ =	3.5553 - $\frac{1}{25}$ (92.5634)=	-0.1472	$\Rightarrow \partial_{3(14)} = e^{-0.1472}$ =	0.8631
$\hat{u}_{3(15)}$ =	3.4340 - $\frac{1}{25}$ (92.5634)=	-0.2685	$\Rightarrow \partial_{3(15)} = e^{-0.2685}$ =	0.7645
$\hat{u}_{3(16)}$ =	3.4012 - $\frac{1}{25}$ (92.5634)=	-0.3013	$\Rightarrow \partial_{3(16)} = e^{-0.3013}$ =	0.7398
$\hat{u}_{3(17)}$ =	3.3322 - $\frac{1}{25}$ (92.5634)=	-0.3703	$\Rightarrow \partial_{3(17)} = e^{-0.3703}$ =	0.6905
$\hat{u}_{3(18)}$ =	3.2581 - $\frac{1}{25}$ (92.5634)=	-0.4444	$\Rightarrow \partial_{3(18)} = e^{-0.4444}$ =	0.6412
$\hat{u}_{3(19)}$ =	3.2189 - $\frac{1}{25}$ (92.5634)=	-0.4837	$\Rightarrow \partial_{3(19)} = e^{-0.4837}$ =	0.6165
$\hat{u}_{3(20)}$ =	2.8332 - $\frac{1}{25}$ (92.5634)=	-0.8693	$\Rightarrow \partial_{3(20)} = e^{-0.8693}$ =	0.4192
$\hat{u}_{3(21)}$ =	2.8332 - $\frac{1}{25}$ (92.5634)=	-0.8693	$\Rightarrow \partial_{3(21)} = e^{-0.8693}$ =	0.4192
$\hat{u}_{3(22)}$ =	2.8332 - $\frac{1}{25}$ (92.5634)=	-0.8693	$\Rightarrow \partial_{3(22)} = e^{-0.8693}$ =	0.4192
$\hat{u}_{3(23)}$ =	2.7081 - $\frac{1}{25}$ (92.5634)=	-0.9945	$\Rightarrow \partial_{3(23)} = e^{-0.9945}$ =	0.3699
$\hat{u}_{3(24)}$ =	2.4849 - $\frac{1}{25}$ (92.5634)=	-1.2176	$\Rightarrow \partial_{3(24)} = e^{-1.2176}$ =	0.2959
$\hat{u}_{3(25)}$ =	2.3979 - $\frac{1}{25}$ (92.5634)=	-1.3046	$\Rightarrow \partial_{3(25)} = e^{-1.3046}$ =	0.2713

Con estos resultados se observa que la más influyente es $\partial_{3(1)} = 5.03$ infarto al miocardio en hombres cuya edad no está por debajo de la edad promedio, mientras que la menos influyente es $\partial_{3(25)} = .2713$ Sida.

Conclusiones

Las tablas de contingencias son una herramienta muy útil para el análisis de información de bases de datos, poblaciones, encuestas, etc., es decir, datos cualitativos, ya que permite clasificar la información en dos variables o más. Al presentarla de esta manera, se permite un análisis rápido y eficaz, ya en la tabla se presentan las frecuencias de cada variable y se agrupa la información de tal forma que es fácil compararla.

Al profundizar en el estudio de las tablas de contingencias, se encuentra a los modelos log-lineales, que pueden perfectamente adaptarse a estas tablas, para lo cual se necesita que estás cumplan con ciertas características, pero si son viables, permite revisar las dependencias o independencias entre las variables que forman la tabla, y de esta manera, proporciona mucho más información que la que se observa a simple vista o mediante estadística descriptiva.

Es muy importante aclarar que de acuerdo a la tabla de contingencias que se presenta y a las dependencias y/o independencias que existan entre sus renglones o columnas, debe elegirse el modelo que mejor se adapte a estas tablas, ya que existen diversos modelos que pueden aplicarse de acuerdo a la información que se requiera. Basados en la teoría estadística deben revisarse las características de cada tabla de contingencias y estudiar las distribuciones que pueden utilizarse en cada caso.

En la aplicación del modelo a un caso práctico, se eligió una cartera de siniestros ocurridos de seguros de vida individual, ya que esta cartera contiene diferentes variables cualitativas que pueden estudiarse con este modelo, como la causa de muerte, el sexo, si es fumador o no fumador, el estado de la República Mexicana al que pertenecen, etc., y la combinación entre estas. Al momento de suscribir un riesgo en la práctica, se analizan todas estas variables de manera conjunta, por lo que la hipótesis inicial es que existe dependencia entre las variables estudiadas del modelo.

Se eligieron las variables de causa de muerte, sexo y edad. Para el caso del modelo de 2x2, en el cual se estudiaron las variables de causa de muerte y sexo, la cartera elegida mostró que un porcentaje muy alto es del sexo masculino, 75% contra 25% del sexo femenino, resultado que está de acuerdo con las bases demográficas existentes, ya que se sabe que el sexo masculino tiene una mayor probabilidad de muerte que el sexo femenino.

Con los resultados obtenidos del modelo, se observa que las causas de muerte que están incrementando la frecuencia en ambos sexos, son precisamente las causas particulares de cada sexo, cáncer mamario y cérvico-uterino para el sexo femenino y cáncer prostático para el sexo masculino, lo que también está de acuerdo con los conocimientos del sector.

En el modelo de 3x3, se trató de contrastar los resultados del modelo de 2x2 con la variable edad, es decir, cómo la edad se relacionaba con la causa de edad y el sexo, ya que podría estar ocurriendo que se estuvieran presentando más muertes en edades menores. La edad promedio de la cartera de vida individual es de 38 años, una edad promedio muy común en el sector asegurador, ya que la gente que normalmente tiene acceso a seguros, cuenta con un ingreso económico de nivel medio y en muchos casos, un trabajo establecido y cierta antigüedad.

En el modelo de 3x3, se observa que el porcentaje mayor de siniestros ocurren en edades mayores a la edad promedio de la cartera, por lo que se concluye que no están ocurriendo siniestros a menores de esta edad por ninguna causa que no sea envenenamientos.

Con estos resultados se puede concluir que la cartera se comporta tal y como se esperaba en una cartera de este estilo, no se presentan desviaciones respecto a edad ni sexo.

Bibliografía

- Christetensen, Ronald; ***Log-Linear Models***; Ed. Springer- Verlag; New York, 5 de abril de 1990.

- Mood, Alexander M. y Graybill, Franklin A.; ***Introducción a la Teoría de la Estadística***; McGraw Hill Book Company, 1955.

- Peña Sánchez de Rivera, Daniel; ***Estadísitica Modelos y Métodos***; Alianza Editorial, 1992.

- Peña Sánchez de Rivera, Daniel; ***Estadísitica Modelos y Métodos 2. Modelos lineales y series temporales***; Alianza Editorial, 1992.

- Mendenhall William, Wackerly Dennis D., Scheaffer Richard L.; ***Estadística Matemática con Aplicaciones***; Grupo Editorial Iberoamérica, 1990.

- Escobar Modesto; ***El Análisis Log-Lineal***; Universidad de Salamanca, México D.F., 1999.