Notas de estudio

Oscar Arturo Bringas López

4 de julio de 2018

Análisis de Datos Categóricos

Las variables categóricas tienen una escala de medición consistente en un conjunto de categorías. Por ejemplo, filosofía partidista a menudo se mide como de izquierda, central o de derecha. Los diagnósticos de cáncer de mama basados en las mamografías usan las categorías normal, benigno, probablemente benigno, sospechoso y maligno. En las siguientes notas estudiaremos los métodos para analizar la información proveniente de estos datos.

1. INTRODUCCIÓN

1.1. Distribuciones e Inferencia para Datos Categóricos

El desarrollo de métodos para variables categóricas fue estimulado por estudios de investigación en las ciencias biomédicas y sociales. Las escalas categóricas fueron penetrando en las ciencias sociales para la **medición de actitudes y opiniones.** Las escalas categóricas en las ciencias biomédicas miden resultados tales como si un tratamiento es exitoso.

A pesar de que los datos categóricos son comunes en las ciencias sociales y biomédicas, no existe ninguna restricción para aplicarlo en otras áreas. La industria, la tecnología, el sector de educación, marketing, entre otros, hacen uso de las técnicas disponibles para este tipo de datos. Las variables categóricas son de distinto tipo. Estudiaremos ahora los modos de clasificar los tipos y la forma en la que se relacionan entre sí.

1.1.1. Distinción entre variables de respuesta y variables explicativas

La mayorías de los análisis estadísticos distinguen entre *variables de respuesta (ó dependientes) y variables explicativas (ó independientes).* Por ejemplo, los modelos de regresión describen qué tanto la media de una variable de respuesta, tal como el precio de venta de una casa, cambia de acuerdo a los valores de las variables explicativas, tales como la localización y los metros cuadrados. En estas notas nos enfocaremos en los métodos para las variables categóricas de respuesta

1.1.2. Distinción en la escala Nominal y Ordinal

Las variables categóricas tienen dos tipos primarios de escalas. Las variables con categorías sin un orden natural son llamadas **nominales**. Por ejemplo: la afiliación religiosa con las categorías (1=Católica, 2=Protestante, 3=Judía, 4=Musulmán, 5=Otras), el tipo de transporte al trabajo (1=automóvil, 2=bicicleta, 3=autobús, 4=metro, 5=metrobús, 6=caminando, 7=Otros),

género favorito de música (1=clásica, 2=jazz, 3=rock, 4=pop, etc). **Para las variables nominales, el orden en el que son listadas sus categorías es irrelevante. El análisis estadístico no depende de tal orden**.

Algunas variables categóricas tienen ordenadas sus categorías. Tales variables son llamadas **ordinales.** Algunos ejemplos son: tamaño de automóvil (1=Subcompacto, 2=Compacto, 3=Mediano, 4=Largo), la clase social (1=Baja, 2=Media, 3=Alta), la condición médica de un paciente (1=Buena, 2=Delicada, 3=Seria, Crítica). Las variables ordinales tienen orden en sus categorías, pero **la distancia entre las categorías es desconocida.** A pesar de que una persona categorizada como moderada en cuanto a su preferencia política sea más liberal que una persona categorizada como conservadora, no existe un valor numérico que describa *cuánto más* liberal es esta persona. Los métodos para variables ordinales usan categorías ordenadas.

Una ***variable intervalo*** es aquella que tiene distancias numéricas entre dos valores. Por ejemplo, nivel de presión sanguínea, tiempo de vida funcional de una televisión, lapso de tiempo en prisión, e ingresos anuales son variables intervalo. Una variable intervalo es llamada algunas veces **variable de razón** (*ratio variable*) si las razones de los valores también son válidos.
Nota: Explicar ejemplo de temperaturas.

El modo en que una variable es medida determina su clasificación. Por ejemplo, "educación" es sólo nominal cuando es medida como escuela pública o privada; es ordinal cuando es medida por el grado más alto alcanzado, usando las categorías: sin escolaridad, primaria, secundaria, medio superior, etc. Es una variable de intervalo cuando es medida por el número de años de educación, usando los enteros 0,1,2,3,...

La escala de medición de una variable determina cuál método estadístico es apropiado. En la medición **jerárquica**, las variables intervalo son las más alta, las variables ordinales son las que siguen y las variables nominales son las de jerarquía más baja. Los métodos estadísticos para variables de un tipo pueden también ser usados con variables de **más alto nivel** pero no con los de menor nivel. Por ejemplo, los métodos estadísticos para variables nominales pueden ser usados con variables ordinales al ignorar el orden de las categorías. Los métodos para variables ordinales sin embargo no pueden ser usados con variables nominales, pues sus categorías no tienen ningún significado en el orden. Usualmente es mejor aplicar los métodos apropiados para la escala correspondiente.

Como nuestro estudio se enfoca en las respuestas categóricas, discutiremos con mayor énfasis las variables nominales y ordinales. Los métodos también aplican a las variables de intervalo que tienen un número pequeño de valores distintos (e.g., número de veces que un individuo se ha casado) o para los valores que han sido agrupados en categorías ordenadas (e.g., años de educación como < 10 años, De 10 a 12 años, > 12 años).

1.1.3. Distinción entre variable Continua y Discreta

Las variables son clasificadas como continuas o discreta, de acuerdo con el número de valores que pueden tomar. Las mediciones actuales de todas las variables ocurren de una forma discreta, debido a las limitaciones de precisión en los instrumentos de medición. La clasificación continua-discreta, en la práctica, distingue entre variables que toman muchos valores y variables que toman pocos valores. Por ejemplo, los estadistas suelen tratar variables de intervalo discreto que tienen un gran número de valores (como las puntuaciones de los exámenes) como continuas, usándolas en métodos para respuestas continuas.

En este estudio trataremos con ciertos tipos de respuestas medidas en forma discreta: (1)variables nominales, (2)variables ordinales, (3)variables intervalo discretas que tienen relativamente pocos valores, y (4)variables continuas agrupadas en un pequeño número de categorías.

1.2. Distribuciones de los Datos Categóricos

El análisis inferencias de datos requiere supuestos sobre los mecanismos aleatorios que generan la información. Para los modelos de regresión con variables de respuesta continuas, la distribución normal juega un rol central. En esta sección revisaremos las distribuciones clave para respuestas categóricas: binomial, multinomial y poisson.

1.2.1. Distribución Binomial

Algunas aplicaciones hacen referencia a un número "n"fijo de observaciones binarias. denotemos como $y_1, y_2, ..., y_n$ las respuestas para n ensayos idénticos e independientes tales que $P(Y_i = 1) = \pi$ y $P(Y_i = 0) = 1 - \pi$. Usaremos los nombres "éxito" y "fracaso" para los resultados 1 y 0. *Ensayos idénticos* significa que la probabilidad de éxito π es la misma para cada ensayo. *Ensayos independientes* significa que $\{Y_i\}$ son variables aleatorias independientes. Estos suelen ser llamados *ensayos Bernoulli*. El número total de éxitos, $Y = \sum_{i=1}^{n} Y_i$, tiene una *distribución binomial* con parámetros n y π, denotados por bin(n,π).

$$\mathbb{P}(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 0, 1, 2, ..., n \qquad (1.1)$$

Donde el coeficiente binomial $\binom{n}{y} = \frac{n!}{y!\,(n-y)!}$. Como $\mathbb{E}(Y_i) = \mathbb{E}(Y_i^2) = 1 \cdot \pi + 0 \cdot (1-\pi) = \pi$,

$$\mathbb{E}(Y_i) = \pi \quad y \quad \mathbb{V}ar(Y_i) = \pi(1-\pi).$$

La distribución binomial $Y = \sum_{i=1}^{n} Y_i$ tiene media y varianza

$$\mu = E(Y) = n\pi \quad y \quad \sigma^2 = Var(Y) = n\pi(1-\pi).$$

La distribución converge a una Normal en la medida en que n incrementa, para π fija.

No existe ninguna garantía de que las sucesivas observaciones binarias sean independientes o idénticas. Así que, ocasionalmente utilizaremos otras distribuciones y analizaremos casos para los cuales los supuestos no se cumplen.

1.2.2. Distribución Multinomial

Algunos ensayos tienen más de dos posibles resultados. Supongamos que cada uno de los n independientes e idénticos ensayos puede resultar en cualquiera de c categorías. Denotemos a $y_{ij} = 1$ si el ensayo i cayó en la categoría j y $y_{ij} = 0$ en cualquier otro caso. Entonces y_i=$(y_{i1}, y_{i2}, ..., y_{ic})$ representa un ensayo multinomial, con $\sum_j y_{ij} = 1$; por ejemplo, (0,0,1,0) denota un resultado de la categoría 3 de cuatro posibles categorías. Notemos que y_{ic} es redundante, siendo linealmente dependiente de los otros. Denotemos $n_j = \sum_i y_{ij}$ como el número de ensayos resultantes en la categoría j. Los conteos $(n_1, n_2, ..., n_c)$ tienen una *distribución multinomial*.

Denotemos como $\pi_j = P(Y_{ij} = 1)$ a la probabilidad de que el resultado en la categoría j para cada ensayo. La función de densidad de probabilidad multinomial es

$$P(n_1, n_2, ..., n_{c-1}) = \left(\frac{n!}{n_1! n_2! \cdots n_c!}\right) {\pi_1}^{n_1} {\pi_2}^{n_2} \cdots {\pi_c}^{n_c}. \qquad (1.2)$$

Como $\sum_j n_j = n$, esta es (c-1)-dimensional, con $n_c = n - (n_1 + \cdots + n_{c-1})$. La distribución binomial es el caso especial para el que c=2. Para la distribución multinomial,

$$E(n_j) = n\pi, \qquad Var(n_j) = n\pi_j(1 - \pi_j), \qquad Cov(n_j, n_k) = -n\pi_j\pi_k. \qquad (1.3)$$

Nota: La distribución marginal para cada n_j tiene una distribución binomial.

1.2.3. Distribución Poisson

Algunas veces, los conteos de información no resultan de un número fijo de ensayos. Por ejemplo, si y = número de muertes debido a accidentes de automóvil en autopistas de Italia durante la semana en curso, no existe un límite superior fijo n para y. Como y debe ser un entero no negativo, su distribución debe situarse en cierto rango. La distribución más simple es la distribución *Poisson*. Su probabilidad depende de un único parámetro, la media μ. La función de masa de probabilidad es:

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, ... \qquad (1.4)$$

Se satisface que E(Y)=Var(Y)=μ. Es unimodal con moda igual a la parte entera de μ. La distribución se aproxima a una Normal en la medida en que μ incrementa.

La distribución Poisson es usada parea conteo de eventos que ocurren aleatoriamente en el tiempo o espacio, cuando los resultados de periodos o regiones disjuntos son independientes. También se aplica como una aproximación para la Binomial cuando n es grande y π es pequeño. con $\mu = n\pi$. Así que si cada una de las 50 millones de personas manejando en Italia la próxima semana es un ensayo independiente con probabilidad 0.000002 de morir en un accidente fatal en esa semana, el número de muertes $\mathbf{Y}$ es una variable Bin(50000000,0.000002), o aproximadamente Poisson con $\mu = n\pi = 50000000(0.000002) = 100$.

Una característica clave de la distribución Poisson es que su varianza iguala a su media. Los conteos muestrales varían más cuando su media es más grande. Cuando el número medio de accidentes fatales semanales iguala a 100, ocurre que la variabilidad en los conteos semanales es más grande que cuando la media es igual a 10.

1.2.4. Sobredispersión

En la práctica, las observaciones de conteos suelen exhibir una variabilidad excedente a la predicha por la Binomial o Poisson. Este fenómeno es llamado *sobredispersión*. Asumimos en lo anterior que cada persona tiene la misma probabilidad de muerte en un accidente fatal en la siguiente semana. De forma más realista, estas probabilidades varían debido a factores tales como la cantidad de tiempo invertido manejando, si la persona usa el cinturón de seguridad, y la localización geográfica. Tal variación de la mortalidad cuenta para mostrar más variación de lo previsto por el modelo Poisson.

1.2.5. Conexión entre las Distribuciones Poisson y Multinomial

En Italia esta semana, sea y_1 = número de personas que murieron en accidentes de automóvil, y_2 = número de personas que murieron en accidentes de avión, y y_3 = número de personas que murieron en tren. Un modelo Poisson para (Y_1, Y_2, Y_3) trata a éstas como variables aleatorias independientes Poisson, con parámetros (μ_1, μ_2, μ_3). La función de masa de probabilidad conjunta para $\{Y_i\}$ es el producto de estas tres funciones de masa de probabilidad de la forma (1.4) El total $n = \sum Y_i$ también tiene una distribución Poisson, con parámetros $\sum \mu_i$.
Con un muestreo Poisson el conteo total n es aleatorio en vez de fijo. Si asumimos un modelo Poisson pero condicionado a n, $\{Y_i\}$ no mantiene distribuciones Poisson, puesto que cada Y_i no puede exceder n. Dada n, $\{Y_i\}$ tampoco es ya independiente, pues el valor de uno afecta el posible rango de los otros.
Para c variables independientes Poisson, con $E(Y_i) = \mu_i$, se derivará su distribución condicional dada una $\sum Y_i = n$. La probabilidad condicional de un conjunto de conteos $\{n_i\}$ que satisface esta condición es

$$P\left[(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c) | \sum_j Y_j = n\right] =$$

$$\frac{P(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c)}{P(\sum Y_j = n)} =$$

$$\frac{\prod_i [\frac{e^{-\mu_i} {\mu_i}^{n_i}}{n_i!}]}{\frac{e^{(-\sum \mu_j)} (\sum \mu_j)^n}{n!}} = \frac{n!}{\prod_i n_i!} \prod_i {\pi_i}^{n_i} \qquad (1.5)$$

Donde $\{\pi_i = \mu_i / \sum \mu_j\}$. Ésta es la distribución Multinomial $(n, \{\pi_i\})$. Varios análisis de datos categóricos asumen una distribución multinomial. Tales análisis usualmente tienen las mismas estimaciones de parámetros que aquellos análisis que asumen una distribución Poisson, debido a la similaridad en las funciones de verosimilitud.

1.3. Distribución Binomial Negativa

Una distribución que puede usarse como alternativa a una Poisson es la Binomial Negativa. Dado que su varianza es más grande que su media, constituye una excelente alternativa para modelar datos de conteo sobredispersos, que son muy comunes en aplicaciones reales.

La forma estándar de concebir esta distribución es en una situación de *muestreo por cuota*. Este esquema de muestreo es típico de investigaciones de mercado, en las que se pide a un individuo entrevistar a un número no definido de sujetos (n) hasta que una parte de ellos (m: cuota fija) haya contestado afirmativamente a alguna pregunta o haya preferido un producto bajo investigación para su comercialización. Por supuesto, asumimos que la probabilidad, p, de que obtengamos una respuesta afirmativa, es la misma para cualquier sujeto. Este esquema es similar al que se modelaría con una Binomial, pero, mientras en la Binomial el número de "éxitos", m, es aleatorio y el número de ensayos, n, es fijo, en este caso sucede exactamente al revés: el número de "éxitos", m, es fijo (no aleatorio) y el número de ensayos (n: total de entrevistados necesarios para tener m éxitos) es aleatorio.

Para deducir la función de masa de probabilidad, consideremos el número de fracasos que han ocurrido hasta obtener m éxitos. Si suponemos que han ocurrido k fracasos antes de obtener estos m éxitos, es claro que n=m+k y que en los primeros m+k-1 ensayos debieron de haber ocurrido m-1 éxitos, ya que en el siguiente ensayo debió ocurrir el último éxito que completa

la cuota requerida. Entonces, dado que hemos fijado la condición de m-1 éxitos en los primeros m+k-1 ensayos, este hecho puede modelarse a través de una v.a. Binomial(m+k-1, p), y como requerimos que en el último ensayo, que es independiente de los primeros m+k-1, ocurra necesariamente un éxito, entonces la función de densidad de esta v.a. es

$$P(N=k) = \binom{m+k-1}{m-1} \left(\tfrac{\beta}{1+\beta}\right)^{m-1} \left(\tfrac{1}{1+\beta}\right)^{k} * \left(\tfrac{\beta}{1+\beta}\right), \quad entonces$$

$$P(N=k) = \binom{m+k-1}{k} \left(\tfrac{\beta}{1+\beta}\right)^{m} \left(\tfrac{1}{1+\beta}\right)^{k} \quad k = 0, 1, 2, 3, ...$$

ya que

$$\binom{m+k-1}{m-1} = \binom{m+k-1}{k}$$

con $p = \dfrac{\beta}{1+\beta}$ la probabilidad de éxito.
La distribución Binomial Negativa tiene media y varianza

$$E(N) = k\beta; \quad V(N) = k\beta(1+\beta)$$

Ya que $\beta > 0$, entonces la varianza de la binomial negativa es mayor que su media, razón por la que suele usarse como alternativa a la Poisson cuando ésta es sobredispersa (Var ¿media).

1.4. Distribución Hipergeométrica

Supongamos que tenemos una población dividida en dos sub-poblaciones excluyentes y exhaustivas, de tamaño n1 y n2 , respectivamente (n = n1 + n2) y seleccionamos k individuos de los n totales. Si nos preguntamos por la probabilidad de que **z"** de estos individuos seleccionados pertenezcan a la primera sub-población, la variable aleatoria que modela este tipo de fenómenos, recibe el nombre de Hipergeométrica. La manera de construir la función de densidad de probabilidad asociada a esta variable es

1. Considerar la selección de k individuos de n posibles

$$\binom{n}{k} \quad "Casos \quad totales"$$

2. Seleccionar y individuos de la primera sub-población

$$\binom{n_1}{y}$$

3. Los otros k-y, necesariamente deben pertenecer a la sub-población 2, y los elegimos de la siguiente manera

$$\binom{n_2}{k-y}$$

Finalmente, la probabilidad de elegir y individuos de la sub-población 1 es

$$P(Y=y)=\frac{\binom{n_1}{y}\binom{n_2}{k-y}}{\binom{n}{k}}=\frac{"Casos favorables"}{"Casos totales"}, \qquad y\leq k, \quad y\leq n_1, \quad k-y\leq n_2$$

$$Con \quad E(Y)=k\frac{n_1}{n} \quad y \quad V(Y)=\frac{n_1 n_2 k(n-k)}{n^2(n-1)}$$

La elección de la distribución para la variable de respuesta es sólo un paso en el análisis de datos. En la práctica, la distribución tiene valores de parámetros desconocidos. En ese estudio revisaremos los métodos para uso de información muestral para hacer inferencia sobre los parámetros.

1.5. Inferencia Estadística para Datos Categóricos

1.5.1. Funciones de Verosimilitud y Estimaciones de Máxima Verosimilitud

En nuestro estudio de análisis de datos categóricos, usamos la *máxima verosimilitud* para la estimación de los parámetros. Bajo condiciones de regularidad débil, tal como el espacio parametral con dimensiones fijas y con valor verdadero que cae en su interior, los estimadores máximo verosímil tienen propiedades deseables: Tienen distribuciones normales de gran muestra, son asintóticamente equivalentes, convergen al parámetro en la medida en que n incrementa y son además asintóticamente eficientes, produciendo errores estándar de gran muestra no más grandes que aquellos calculados bajo otros métodos de estimación.

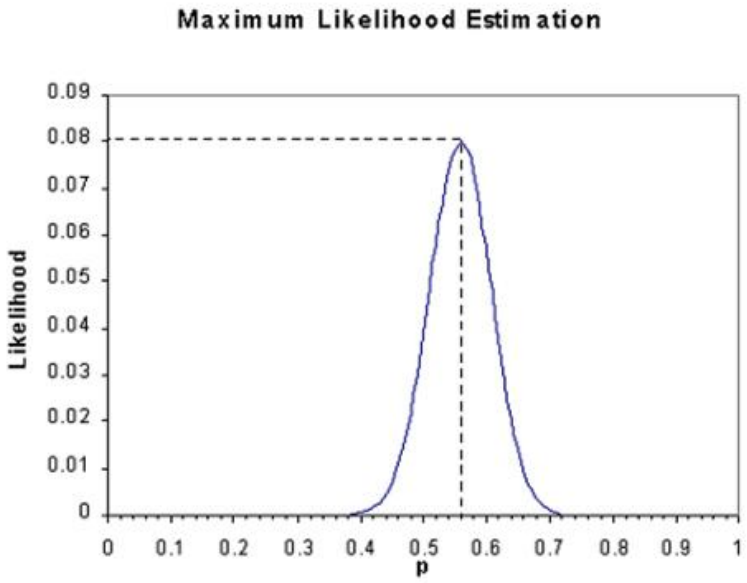
Función de Verosimilitud Sea $X_1, ..., X_n$ una muestra aleatoria (m.a) de $f(X;\theta)$ con θ en el espacio parametral. Dados los valores de la m.a. $X_1, ..., X_n$ se define a la función de verosimilitud como la función de densidad muestral evaluada en $(X_1, X_2, ...X_n)$ y vista como función de θ. Esto es,

$$L(\theta)=\prod_{i=1}^{n} f(X_i;\theta)$$

Estimador Máximo Verosímil, $\hat{\theta}_{MV}$: Dada $X_1, ..., X_n$ m.a de $f(X;\theta)$, se define al estimador máximo verosímil, como aquél que satisface:

$$L(\hat{\theta}_{MV})=sup_{\theta\in H}L(\theta)$$

El estimador máximo verosímil es el valor paramétrico que maximiza esta función. Éste es el valor bajo el cual la información observada tiene la probabilidad más alta de ocurrencia. Deseamos encontrar el valor de θ que maximice $L(\theta)$. Al ser función continua y diferenciable entonces podemos encontrar el máximo con criterio de la derivada y evaluando en puntos extremos.



Deberíamos obtener $\frac{\partial}{\partial\theta}L(\theta)$ y con esta obtener el máximo, sin embargo, esta tarea puede simplificarse si encontramos el máximo de la función

$$\ell(\theta) := \ln\left[L(\theta)\right] \qquad \text{“Función de log-verosimilitud”}$$

Nota: Al ser $\ln(\cdot)$ una función creciente, entonces $\ell(\theta)$ alcanza el máximo en el mismo punto que $L(\theta)$.

Para varios modelos, $\ell(\theta)$ tiene una forma cóncava y $\hat{\theta}$ es el punto en el cuál la derivada es igual a 0. La estimación máximo verosímil es entonces la solución a la ecuación de verosimilitud, $\partial\ell(\theta)/\partial\theta = 0$. A menudo, θ es multidimensional, denotado por $\underline{\theta}$ y $\hat{\underline{\theta}}$ es la solución del conjunto de ecuaciones de verosimilitud.

Función Score: A la función $\frac{\partial}{\partial\theta}\ln\left(f(\underline{X};\theta)\right)$ se le llama función de puntaje o *función score*

Información Esperada de Fisher: A la cantidad $I_x(\theta) = \mathbb{E}[(\frac{\partial}{\partial\theta}\ln f(X_i;\theta))^2]$ se le conoce como *Información Esperada de Fisher por unidad muestral* y representa intuitivamente la “cantidad de información” acerca del valor del parámetro θ contenida en una observación de la v.a. X

Nota importante: Se demuestra en el curso de Estadística I que:

- $$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_i;\theta)\right] = I_x(\theta)$$
- $$I_{\underline{x}}(\theta) := nI_x(\theta)$$

Cuando tenemos parámetros multidimensionales por estimar, es posible construir una matriz de información, conformada por cada una de las funciones de información esperada de Fisher.

Denotemos por **SE** al *error estándar* de $\hat{\theta}$, y denotemos con $cov(\underline{\hat{\theta}})$ a la matriz de covarianzas asintóticas de $\underline{\hat{\theta}}$. **Bajo condiciones regulares**, $cov(\underline{\hat{\theta}})$ es la inversa de la matriz de información. El (j,k) elemento de la matriz de información es

$$-\mathbb{E}\left[\frac{\partial^2 \ell\left(\underline{\theta}\right)}{\partial\theta_j \partial\theta_k}\right]$$

Los errores estándar son la raíz cuadrada de los elementos de la diagonal de la inversa de la matriz de información, ya que como se vio en el curso de estadística I, haciendo uso de la teoría de la Cota Inferior de Cramer Rao (CICR)

$$Var(T(\underline{X})) \geq \frac{1}{I_{\underline{X}}(\theta)} = \frac{1}{nI_X(\theta)}$$

Ejemplo: Función de Verosimilitud y Estimación de Máxima Verosimilitud para un parámetro Binomial

La parte de la función de verosimilitud que involucra los parámetros es llamada *Kernel*. Como la maximización de la verosimilitud es respecto a los parámetros, el resto es irrelevante.

Para ilustrarlo, consideremos la distribución binomial. El coeficiente binomial $\binom{n}{y}$ no tiene influencia en donde ocurre la maximización con respecto a π. Así que lo ignoramos y tratamos con el kernel como la función de verosimilitud. La log-verosimilitud binomial es entonces

$$\ell(\pi) = \ln\left[\pi^y(1-\pi)^{n-y}\right] = y\ln(\pi) + (n-y)\ln(1-\pi)$$

Diferenciando con respecto a π obtenemos

$$\frac{\partial \ell(\pi)}{\partial \pi} = \frac{y}{\pi} - \frac{(n-y)}{(1-\pi)} = \frac{y - n\pi}{\pi(1-\pi)}$$

Igualando ésto a 0 obtenemos la ecuación de verosimilitud, la cual tiene solución $\hat{\pi} = {}^{y}/_{n}$, la proporción muestral de éxitos en n ensayos.

Calculando ${}^{\partial^2 \ell(\pi)}/_{\partial \pi^2}$, tomando su esperanza y combinando términos, obtenemos

$$-\mathbb{E}\left[\frac{\partial^2 \ell(\pi)}{\partial \pi^2}\right] = \mathbb{E}\left[\frac{y}{\pi^2} + \frac{(n-y)}{(1-\pi)^2}\right] = \frac{n}{\pi(1-\pi)}$$

Por lo que la varianza asintótica de $\hat{\pi}$ es $\frac{\pi(1-\pi)}{n}$. Por lo que la distribución de $\hat{\pi} = {}^{Y}/_{n}$ tiene media y error estándar

$$\mathbb{E}(\hat{\pi}) = \pi, \qquad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

1.5.2. Triada de Pruebas Wald-Likelihood Ratio-Score

Existen tres modos estándar de usar la función de verosimilitud para realizar inferencia en grandes muestras. Los introducimos para una prueba de significancia de la hiótesis nula H_0 : $\beta = \beta_0$ y posteriormente discutiremos su relación con la estimación de intervalos de confianza. Todos estos métodos explotan la normalidad de las muestras grandes de los estimadores máximo verosímiles.

Estadístico de Wald Con error estándar no nulo **SE** de $\hat{\beta}$, el estadístico de prueba

$$z = \frac{\hat{\beta} - \beta_0}{SE}$$

se aproxima a una distribución normal estándar cuando $\beta = \beta_0$. Se hace referencia a "z" con la tabla normal estándar para obtener los p-values de una o dos colas. Equivalentemente, para la alternativa de dos colas, z^2 tiene una distribución nula Ji-cuadrada con un agrado de libertad; el p-value es entonces de cola derecha sobre los valores observados. Este tipo de estadística, usando el error estándar no nulo, es llamado **Estadístico de Wald** (Wald 1943).
La extensión multivariada para la prueba de Wald en $H_0 : \underline{\beta} = \underline{\beta}_0$ tienes estadística de prueba

$$W = (\underline{\hat{\beta}} - \underline{\beta}_0)'[cov(\underline{\hat{\beta}})]^{-1}(\underline{\hat{\beta}} - \underline{\beta}_0).$$

(La prima en un vector o matriz denota la función transpuesta.) La covarianza no-nula está basada en la curvatura de la log-verosimilitud de $\underline{\hat{\beta}}$. La distribución normal multivariada asintótica para $\underline{\hat{\beta}}$ implica una distribución Ji-cuadrada asintótica para W. Los grados de libertad igualan al rango de $cov(\underline{\hat{\beta}})$, el cual es el número de parámetros no redundantes en $\underline{\beta}$.

Estadístico Likelihood Ratio También llamado cociente de verosimilitudes. Es un segundo método que usa la función de verosimilitud mediante el cociente de dos maximizaciones: (1) El máximo sobre los posibles valores parametrales bajo H_0, y (2) El máximo sobre el conjunto mayor de valores paramétricos que permite que H_0 o una alternativa H_a sean verdaderas. Denotemos por Ł$_0$ al valor maximizado de la función de verosimilitud bajo H_0, y denotemos por Ł$_1$ al valor maximizado general (i.e., bajo $H_0 \cup H_a$). Ł$_1$ es entonces siempre al menos tan grande como Ł$_0$, pues Ł$_0$ resulta de maximizar sobre un conjunto restringido de valores parametrales.

El cociente $\Lambda = {}^{Ł_0}/_{Ł_1}$ de las verosimilitudes maximizadas no pueden exceder a 1. Wilks (1935,1938) mostró que $-2log\Lambda$ tiene una distribución nula limitante Ji-cuadrada, como n tienda a infinito. Los grados de libertad igualan a la diferencia en las dimensiones del espacio parametral bajo $H_0 \cup H_a$ y bajo H_0. La estadística de prueba *likelihood ratio* (o cociente de verosimilitudes) es igual a

$$-2log\Lambda = -2log({}^{Ł_0}/_{Ł_1})) - 2(\ell_0 - \ell_1),$$

donde $\ell_0 \quad y \quad \ell_1$ denotan a las funciones de máxima log-verosimilitud.

Estadístico Score: El tercer método hace uso de la *estadística Score*. La prueba score está basada en la pendientes y la curvatura esperada de la función de log-verosimilitud $\ell(\beta)$ en el valor nulo β_0. Ésta utiliza el tamaño de la función score

$$u(\beta) = \frac{\partial \ell(\beta)}{\partial \beta},$$

evaluado en β_0. El valor $u(\beta_0)$ tiende a ser más largo en valor absoluto cuando $\hat{\beta}$ está más lejos de β_0. Denotemos a $-\mathbb{E}[\partial^2\ell(\beta)/\partial\beta^2]$ (i.e., la información) evaluada en β_0 por $\imath(\beta_0)$. La estadística score es el cociente de $u(\beta_0)$ y su **SE** nulo, el cual es $[\imath(\beta_0)]^{1/2}$. Ésta tiene una distribución nula aproximada normal estándar. La forma Ji-cuadrada de la estadística score es

$$\frac{[u(\beta_0)]^2}{\imath(\beta_0)} = \frac{[\partial\ell(\beta)/\partial\beta_0]^2}{-\mathbb{E}[\partial^2\ell(\beta)/\partial\beta_0^2]},$$

Donde la notación de la derivada parcial refleja derivadas con respecto a β que son evaluadas en β_0. En el caso multiparamétrico, la estadística score es una forma cuadrática basada en el vector de derivadas parciales de la log-verosimilitud con respecto a $\underline{\beta}$ y la inversa de la matriz de información, ambos evaluados en la estimación H_0 (i.e., asumiendo $\underline{\beta} = \underline{\beta_0}$).

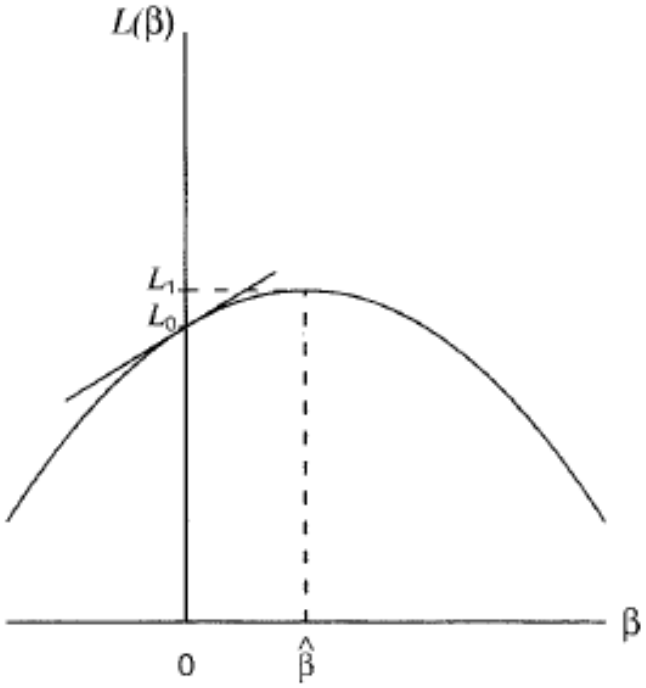


Figura 1: Función de información y verosimilitud usada en las tres pruebas de $H_0 : \beta = 0$

La figura 1 es una gráfica genérica de la log-verosimilitud $\ell(\beta)$ para el caso univariado. Ésta ilustra las tres pruebas de $H_0 : \beta = 0$. **La prueba de Wald** usa el comportamiento de $\ell(\beta)$ en la estimación máximo verosímil β, teniendo una forma ji-cuadrada $(\hat{\beta}/SE)^2$. El error estándar SE de $\hat{\beta}$ depende de la curvatura de $\ell(\beta)$ en $\hat{\beta}$. **La prueba score** está basada en la cuesta abajo y curvatura de $\ell(\beta)$ en $\beta = 0$. **La prueba Likelihood ratio** combina la información sobre $\ell(\beta)$ tanto en $\hat{\beta}$ como en $\beta_0 = 0$. Ésta compara los valores log-verosímiles ℓ_1 en $\hat{\beta}$ y ℓ_1 en $\beta_0 = 0$ usando la estadística ji-cuadrada $-2(\ell_0 - \ell_1)$. En la figura 1, esta estadística es el doble de la distancia vertical entre los valores de $\ell(\beta)$ en $\hat{\beta}$ y en 0. En un sentido, esta estadística usa la mayor información de los tres tipos de pruebas estadísticas y es la más versátil.

En la medida en que $n \to \infty$, las pruebas Wald, Likelihood ratio y score tienen cierta equivalencia asintótica. Para tamaños de muestra pequeñas a moderadas, la prueba likelihood ratio es usualmente más confiable que la prueba Wald.

1.5.3. Construcción de Intervalos de Confianza

En la práctica, es más informativo construir intervalos de confianza para los parámetros que hacer pruebas de hipótesis sobre los valores. Para cualquiera de los tres métodos, el intervalo de confianza resulta de invertir la prueba. Por ejemplo, un intervalo de confianza del 95 % para β es el conjunto de β_0 para los cuales la prueba de $H_0 : \beta = \beta_0$ tiene un p-value excedente a 0.05.

Denotemos por Z_a a la estadística z-score de la distribución normal estándar que tiene probabilidad de cola derecha a; esto es el percentil 100(1-a) de esa distribución. Sea $\mathcal{X}^2_{df}(a)$ quien denote el percentil 100(1-a) de la distribución Ji-cuadrada con grados de libertad df. Los intervalos de confianza $100(1-\alpha)$ % basados en la normalidad asintótica usan $Z_{\alpha/2}$, por ejemplo $Z_{0.025} = 1.96$ para 95 % de confianza.

El intervalo de confianza Wald es el conjunto de β_0 para el cual $\left|\hat{\beta} - \beta_0\right| / SE < Z_{\alpha/2}$. Ésto da el intervalo $\hat{\beta} \pm Z_{\alpha/2}(SE)$.

El intervalo de confianza basado en el cociente de verosimilitudes es el conjunto de β_0 para los cuales $-2\left[\ell(\beta_0) - \ell(\hat{\beta})\right] < \mathcal{X}^2_1(\alpha)$. Recordemos que $\mathcal{X}^2_1(\alpha) = \mathcal{Z}^2_{\alpha/2}$.

Cuando $\hat{\beta}$ tiene una distribución normal, la función de log-verosimilitud tiene una forma parabólica (i.e., un polinomio de segundo grado). Para pequeñas muestras con información categórica, $\hat{\beta}$ puede estar lejos de la normalidad y la función de log-verosimilitud puede estar lejos de una curva con forma parabólica simétrica. Una divergencia marcada en los resultados de inferencia Wald y del cociente de verosimilitudes indican que la distribución de $\hat{\beta}$ puede no estar cerca de la normalidad. El ejemplo que veremos más adelante ilustra lo anterior con intervalos de confianza bastante diferentes para métodos diferentes. En algunos de tales casos, la inferencia puede en cambio utilizar una distribución exacta de muestras pequeñas o métodos asintóticos de orden superior que mejoren la normalidad simple.

El intervalo de confianza Wald es el más común en la práctica debido a que es simple de construir usando estimaciones máximo verosímil y los errores estándar reportados por los softwares estadísticos. El intervalo basado en el cociente de verosimilitudes está siendo más ampliamente disponible en los softwares y es preferible para datos categóricos de pequeñas y moderadas n. Para el mejor modelo estadístico conocido como regresión, para respuestas normales, los tres tipos de inferencia necesariamente proveen resultados idénticos.

Inferencia Estadística para Parámetros Binomiales

En esta sección mostraremos métodos de inferencia para datos categóricos presentando pruebas e intervalos de confianza para los parámetros binomiales π, basados en y éxitos en n ensayos independientes. Anteriormente ya obtuvimos la función de verosimilitud y el estimador M.V. $\hat{\pi} = {}^{y}/_{n} \quad de \quad \pi$.

Pruebas sobre Parámetros Binomiales

Consideremos $H_0 : \pi = \pi_0$. Como H_0 tiene un parámetro único, usamos la forma normal en lugar de ji-cuadrada para las pruebas estadísticas de Wald y Score. Ellas permiten pruebas contra una sola cola, así como las alternativas de dos colas. La estadística de Wald es

$$Z_W = \frac{\hat{\pi} - \pi_0}{SE} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}.$$

Evaluando la información y la función score binomial en π_0 obtenemos

$$u(\pi_0) = \frac{y}{\pi_0} - \frac{n-y}{1-\pi_0}, \qquad \imath(\pi_0) = \frac{n}{\pi_0(1-\pi_0)}.$$

La forma normal del estadístico score se simplifica a

$$Z_s = \frac{u(\pi_0)}{[\imath(\pi_0)]^{1/2}} = \frac{y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}_0(1-\hat{\pi}_0)/n}}.$$

Mientras que el estadístico de Wald Z_W usa el error estándar evaluado en $\hat{\pi}$, el estadístico score Z_S lo usa evaluado en π_0. El estadístico score es preferible, pues usa el erro estándar nulo real en vez de uno estimado. Su distribución muestral nula está más cerca de una normal estándar que la estadística de Wald.

La función Binomial de log-verosimilitud se iguala a $\ell_0 = y\,log\pi_0 + (n-y)\,log(1-\pi_0)$ bajo H_0 y $\ell_1 = y\,log\hat{\pi} + (n-y)\,log(1-\hat{\pi})$ más generalmente. La prueba estadística del cociente de verosimilitudes se simplifica a

$$-2(\ell_0 - \ell_1) = 2\left(y\,log\frac{\hat{\pi}}{\pi_0} + (n-y)\,log\frac{1-\hat{\pi}}{1-\pi_0}\right).$$

Expresado como

$$-2(\ell_0 - \ell_1) = 2\left(y\,log\frac{y}{n\pi_0} + (n-y)\,log\frac{n-y}{n-n\pi_0}\right),$$

Éste compara los conteos de éxitos y fracasos observados con los conteos esperados (i.e., nulo)

$$2\sum observados\ log\frac{observados}{esperados}.$$

Intervalo de Confianza para Parámetro Binomial

Una prueba de significancia indica meramente si un valor π particular (tal como $\pi = 0.05$) es plausible. Aprendemos más usando intervalos de confianza para determinar el rango de valores plausibles.
Invirtiendo la prueba estadística de Wald se obtiene el intervalo del valor π_0 para el cual $|Z_W| < Z_{\alpha/2}$, ó

$$\hat{\pi} \pm\ Z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

Históricamente, éste fue uno de los primero intervalos de confianza usados para cualquier parámetro. Desafortunadamente, se realiza pobremente a no ser que n sea demasiado grande. La probabilidad de cobertura real generalmente cae por debajo del coeficiente de confianza nominal, muy por debajo cuando π está cerca de 0 o 1. Un simple ajuste que adhiere $\frac{1}{2}z^2_{\alpha/2}$ observaciones de cada tipo a la muestra antes de usar esta fórmula funciona mucho mejor.
El intervalo de confianza Score contiene valores π_0 para los cuales $|z_s < z_{\alpha/2}|$. Sus puntos finales son las soluciones π_0 a la ecuaciones

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \pm\, z_{\alpha/2}.$$

Éste es cuadrático en π_0. Por primera vez discutido por E. B. Wilson (1927), este intervalo es

$$\hat{\pi}\left(\frac{n}{n+z_{\alpha/2}^2}\right)+\frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}^2}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+z_{\alpha/2}^2}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)\right]}.$$

El punto medio $\tilde{\pi}$ del intervalo es un promedio ponderado de $\hat{\pi}$ y $\frac{1}{2}$, donde el peso $n/(n+z_{\alpha/2}^2)$ dado a $\hat{\pi}$ incrementa como n incremente. Combinando términos, este punto medio iguala $\tilde{\pi} = (y+\frac{z_{\alpha/2}^2}{2})/(n+z_{\alpha/2}^2)$. Ésta es la proporción muestral para una muestra ajustada que adhiere $z_{\alpha/2}^2$ observaciones, la mitad a cada tipo. El cuadrado del coeficiente $z_{\alpha/2}$ en esta fórmula es un promedio ponderado de la varianza de la proporción muestral cuando $\pi = \frac{1}{2}$, usando el tamaño de muestra ajustado $n+z_{\alpha/2}^2$ en lugar de n. Este intervalo tiene mucho mejor rendimiento que el intervalo de Wald.

El intervalo de confianza basado en el cociente de verosimilitudes es mucho más complejo computacionalmente, pero simple en principio. Éste es el conjunto de π_0 para el cual la prueba de cociente de verosimilitud tiene un *P-value* excedente a α. Equivalentemente, éste es el conjunto de π_0 para el cual el doble de la log-verosimilitud cae por debajo de $\mathcal{X}_1^2(\alpha)$ de su valor en la estimación M.V. $\hat{\pi} = y/n$.

Ejemplo de la proporción de Vegetarianos

En un salón de clases se hizo un cuestionario, se le preguntó a cada estudiante si él o ella eran vegetarianos. De n=25 estudiantes, y=0 contestaron "sí". Ellos no fueron una muestra aleatoria de una población en particular, pero usamos esta información para ilustrar el intervalo de confianza al 95 % para un parámetro binomial π.
Como y=0, $\hat{\pi} = 0/25$. Usando la aproximación de Wald, el intervalo de confianza 95 % para π es

$$0 \pm 1.96\sqrt{\frac{0.00\,x\,1.00}{25}}, \qquad ó \qquad (0,0).$$

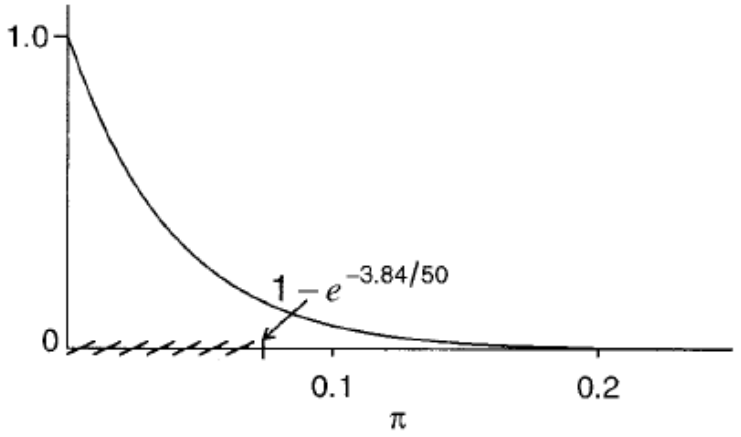
Cuando la observación cae en el límite del espacio muestral, a menudo los métodos de Wald no proporcionan respuestas sensatas.
En contraste, el intervalo de confianza score 95 % iguala (0.0,0.133). Ésta es una inferencia más creíble. Para $H_0 : \pi = 0.5$, por ejemplo, la estadística de prueba score es $z_s = (0-0.5)/\sqrt{(0.5\,x\,0.5)/25} = -5.0$, así que 0.5 no cae en el intervalo.
Por el contrario, para $H_0 : \pi = 0.10$, $\quad z_s = (0-0.10)/\sqrt{(0.10\,x\,0.90)/25} = -1.67$, así que 0.10 cae en el intervalo.

Cuando y=0 y n=25, el kernel de la función de verosimilitud es $Ł(\pi) = \pi_0(1-\pi)^{25}$. La log-verosimilitud es $\ell(\pi) = 25\,\ln(1-\pi)$. Notemos que $\ell(\hat{\pi}) = \ell(0) = 0$. El intervalo de confianza 95 % del cociente de verosimilitudes es el conjunto de π_0 para el cual la estadística de cociente de verosimilitud

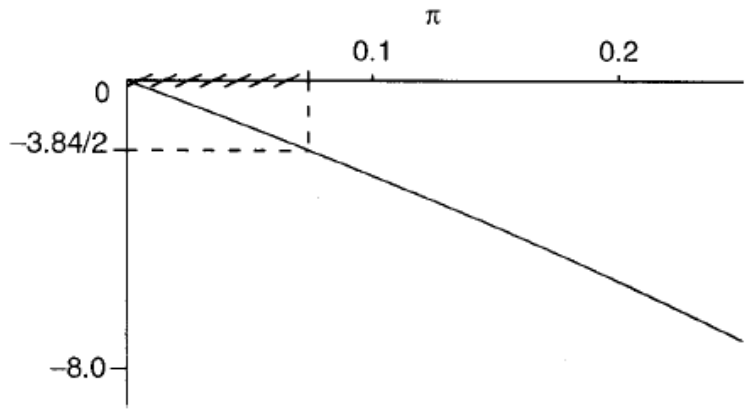
$$\begin{aligned} -2(\ell_0 - \ell_1) &= -2[\ell(\pi_0) - \ell(\hat{\pi})] \\ &= -50\,\ln(1-\pi_0) \leq \mathcal{X}_1^2(0.05) = 3.84. \end{aligned}$$

La cota superior es $1 - e^{-3.84/50} = 0.074$, y el intervalo de confianza iguala (0.0,0.074). La figura 2 muestra las funciones de verosimilitud y log-verosimilitud y su correspondiente región de confianza para π.

Los tres métodos de gran muestra producen resultados muy diferentes. Cuando π es cercano a 0, la distribución muestral de $\hat{\pi}$ es altamente sesgada a la derecha para n pequeñas. Vale la pena considerar métodos alternativos que no requieran aproximaciones asintóticas.



Verosimilitud L(π)



Log-verosimilitud $\ell(\pi)$

Figura 2: Verosimilitud y log-verosimilitud Binomiales cuando y=0 en n=25 ensayos, e intervalo de confianza para π

Inferencia estadística para parámetros multinomiales

Presentamos ahora la inferencia para parámetros multinomiales $\{\pi_j\}$. En n observaciones, n_j ocurren en la categoría j, $j = 1, ..., c$.

Estimación de parámetros Multinomiales

Primero, obtenemos las estimaciones máximo verosímil de $\{\pi_j\}$. Como una función de $\{\pi_j\}$, la función de masa de probabilidad multinomial es proporcional al kernel

$$\prod_j \pi_j^{n_j}, \quad donde \quad todos \quad los \quad \pi_j \geq 0 \quad y \quad \sum_j \pi_j = 1. \tag{1}$$

Las estimaciones M.V. son los $\{\pi_j\}$ que maximizan (1). La función de log-verosimilitud es

$$\ell(\underline{\pi}) = \sum_j n_j \ln \pi_j$$

Para eliminar redundancias, tratamos a ℓ como función de $(\pi_1, ..., \pi_{c-1})$, como $\pi_c = 1 - (\pi_1 + ... + \pi_{c-1})$. Así que, ${}^{\partial \pi_c}/_{\partial \pi_j} = -1$, $j = 1, ..., c-1$.
Como

$$\frac{\partial \ln \pi_c}{\partial \pi_j} = \frac{1}{\pi_c} \frac{\partial \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c},$$

Diferenciando $\ell(\underline{\pi})$ con respecto a π_j obtenemos la ecuación de verosimilitud

$$\frac{\partial \ell(\underline{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0.$$

La solución MV satisface $sfrac\hat{\pi}_j\hat{\pi}_c = {}^{n_j}/_{n_c}$. Ahora

$$\sum_j \hat{\pi}_j = 1 = \frac{\hat{\pi}_c \left(\sum_j n_j \right)}{n_c} = \frac{\hat{\pi}_c \, n}{n_c},$$

así $\hat{\pi}_c = {}^{n_c}/_n$ y entonces $\hat{\pi}_j = {}^{n_j}/_n$. Estos resultados maximizan como veremos más adelante la verosimilitud. Así que, las estimaciones M.V. de $\{\pi_j\}$ son las proporciones muestrales.

1.5.4. Cociente de Verosimilitudes Ji Cuadrada

Una prueba alternativa para los parámetros multinomiales usa la prueba de cociente de verosimilitudes. Vimos anteriormente cuál es el kernel de esta distribución. Bajo H_0 la verosimilitud es maximizada cuando $\hat{\pi}_j = \pi_{j0}$. En el caso general, se maximiza cuando $\hat{\pi}_j = {}^{n_j}/_n$. El cociente de la verosimilitud iguala

$$\Lambda = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j ({}^{n_j}/_n)^{n_j}}.$$

Así que el estadístico de cociente de verosimilitudes, denotado por G^2, es

$$G^2 = -2 \log \Lambda = 2 \sum n_j \log (n_j / n \, \pi_{j0}).$$

Este estadístico, es llamado *estadístico cociente de verosimilitudes Ji cuadrado (likelihood-ratio chi-squared statistic).* Para n grandes, G^2 tiene una distribución nula ji cuadrada con gl=c-1. Cuando H_0 se mantiene, la estadística $\mathcal{X}^2$ de Pearson y el cociente de verosimilitudes G^2 tienen ambos una distribución asintótica ji cuadrada con gl=c-1.

2. TABLAS DE CONTINGENCIA

2.1. Estructura de Probabilidad de las Tablas de Contingencia

Cuando se tienen dos o más variables categóricas, es común desplegar de manera conjunta las relaciones que dan entre ellas, a través de las llamadas tablas de contingencia o tablas de clasificación cruzada. La distribución conjunta de estas dos variables categóricas determina esta relación. Además de determinar sus distribuciones marginales y condicionales.

2.1.1. Tablas de Contingencia y sus Distribuciones

Supongamos que tenemos dos variables categóricas X y Y, X con I categorías y Y con J categorías. Entonces, una clasificación de las respuestas de estas variables tiene IJ posibles combinaciones. Las respuestas (X,Y) de un sujeto seleccionado aleatoriamente de alguna población tiene una distribución de probabilidad asociada. Es posible desplegar esta distribución en una tabla rectangular con I renglones para las categorías de X y J columnas para las de Y. Las celdas de esta tabla representan las IJ posibles combinaciones de respuesta. Cuando las celdas contienen frecuencias de conteos muestrales de respuestas, la tabla se conoce como tabla de contingencia, término introducido por Karl Pearson (1904). Otra forma de denominarla es tabla de clasificación cruzada. Una tabla de contingencia con I renglones y J columnas se conoce como una tabla **IxJ**, o como tabla de contingencia de **IxJ**.

Modelos Muestrales para Tablas IxJ Vimos que existen diversos modelos (distribuciones) para analizar datos categóricos, en este aparatado, estudiaremos las distribuciones que se desprenden al considerar ciertas estructuras de los elementos que componen una tabla de contingencia, que, esencialmente, se obtienen condicionando específicamente sobre algunos de ellos. El estudio lo haremos únicamente con tablas 2 x 2, ya que su extensión a tablas más generales I x J es similar. Introduzcamos la notación que vamos a utilizar en estas tablas. En este caso I = J = 2, por lo que la tabla se puede escribir como

	Y		
X	y_1	y_2	TOTAL
x_1	n_{11}	n_{12}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$n_{2\bullet}$
TOTAL	$n_{\bullet 1}$	$n_{\bullet 2}$	$n = n_{\bullet\bullet}$

Tabla 1: Estructura de Tabla de Contingencia

Entonces, consideremos esta tabla 2x2, con N_{ij} *i,j* = 1,2 la variable aleatoria que denota el número de observaciones (conteos) en la celda (i,j) y un modelo de muestreo *Poisson independiente*

$$N_{ij} = Poisson(\lambda_{ij})$$

Esta es la forma más general de distribución asociada a una tabla de contingencia. Ni marginales ni totales de la tabla se consideran fijos por diseño, y cada celda se considera como una variable aleatoria Poisson independiente. Inicialmente, estudiaremos cuál es la distribución asociada a esta tabla en el caso de que el total de sujetos muestreados se considere fijo, o bien,

condicionando la distribución a un total fijo $N_{\bullet\bullet} = n_{\bullet\bullet}$, y considerando que los marginales varían libremente. En este caso, asumimos que se recaba una muestra de tamaño fijo n, y, posteriormente, las observaciones se clasifican de acuerdo a las categorías de interés (diseño transversal). Primero, observemos que

$$N_{\bullet\bullet} = \sum_{i=1}^{2}\sum_{j=1}^{2} N_{ij} \sim Poisson(\lambda_{ij}), \qquad con \quad \lambda_{\bullet\bullet} = \lambda_{11} + \lambda_{12} + \lambda_{21} + \lambda_{22}$$

Entonces, condicionando por el gran total $N_{\bullet\bullet} = n_{\bullet\bullet}$, la distribución del resto de las tres entradas (recordar que fijo el gran total, el vector de variables aleatorias, $N_{ij}s$, reduce a dimensión tres).

$$\begin{aligned}
f_{N_{11},N_{12},N_{21}|N_{\bullet\bullet}}(n_{11}, n_{12}, n_{21}|n_{\bullet\bullet}) &= \frac{\mathbb{P}[N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{\bullet\bullet} = n_{\bullet\bullet}]}{\mathbb{P}[N_{\bullet\bullet} = n_{\bullet\bullet}]} \\
&= \frac{P[N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}]}{P[N_{\bullet\bullet} = n_{\bullet\bullet}]} \\
&= \frac{\frac{e^{-\lambda_{11}}\lambda_{11}^{n_{11}}}{n_{11}!}\,\frac{e^{-\lambda_{12}}\lambda_{12}^{n_{12}}}{n_{12}!}\,\frac{e^{-\lambda_{21}}\lambda_{21}^{n_{21}}}{n_{21}!}\,\frac{e^{-\lambda_{22}}\lambda_{22}^{n_{22}}}{n_{22}!}}{\frac{e^{-\lambda_{\bullet\bullet}}\lambda_{\bullet\bullet}^{n_{\bullet\bullet}}}{n_{\bullet\bullet}!}} \\
&= \frac{n_{\bullet\bullet}!}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}\pi_{11}^{n_{11}}\,\pi_{12}^{n_{12}}\,\pi_{21}^{n_{21}}\,\pi_{22}^{n_{22}}
\end{aligned}$$

$$Con \quad \pi_{ij} = \frac{\lambda_{ij}}{\lambda_{\bullet\bullet}} \quad , \quad \frac{n_{\bullet\bullet}!}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!} = \binom{n_{\bullet\bullet}}{n_{11}\,n_{12}\,n_{21}\,n_{22}} \quad y \quad \lambda_{\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{ij}.$$

Entonces, condicional a $N_{\bullet\bullet} = n_{\bullet\bullet}$ y generalizando a tablas IxJ, obtenemos que

$$N_{11},\, N_{12}\,,\, \ldots\,, N_{IJ}\,|\,N_{\bullet\bullet} = n_{\bullet\bullet} \;\sim\; Multinomial\;(n_{\bullet\bullet}\,,\,\pi_{11}\,,\,\pi_{12}\,,\,\ldots\,,\,,\,\pi_{IJ})$$

2.1.2. Total de renglón o columnas fijos

Algunas veces los totales marginales de X o Y son fijos, como en un estudio de casos y controles, en el que, por ejemplo, un número fijo de individuos que presentan algún padecimiento, e.g., cáncer pulmonar (los casos) y un número fijo de individuos que no lo presentan (los controles) se muestrean, y después, se compara algún factor de exposición (e.g., condición de fumador) entre estos casos y estos controles. Otro ejemplo puede ser un ensayo clínico en donde el número de sujetos que reciben, digamos, los tratamientos A y B son ambos fijos.
Primeramente, si condicionamos por el total por columna $N_{\bullet 1} = n_{\bullet 1}, N_{\bullet 2} = n_{\bullet 2}$ fijos, obtendremos una nueva distribución para cada una de las entradas de esa columna

Reto: Deducir la distribución resultante de cada una de las celdas al condicionar por el total de los renglones o por el total de las columnas para el caso de tablas 2X2 y en general para tablas IxJ

2.1.3. Más Distribuciones Condicionales

Ahora veamos cuál es la distribución subyacente a la tabla cuando los marginales son fijos, esto es, los renglones $N_{i\bullet} = \sum_{j=1}^{2} N_{ij}$, las columnas $N_{\bullet j} = \sum_{i=1}^{2} N_{ij}$ y por consiguiente el total $N_{\bullet\bullet} = n_{\bullet\bullet}$. Teniendo en mente el supuesto de que las celdas tienen una distribución Poisson $N_{ij} \sim Poisson(\lambda_{ij})$, definamos las siguientes variables:

$$\begin{array}{rl} Y_1 = N_{11} & \\ Y_2 = N_{11} + N_{12} & \text{(fijo)} \\ Y_3 = N_{11} + N_{21} & \text{(fijo)} \\ Y_4 = N_{11} + N_{12} + N_{21} + N_{22} & \text{(fijo)} \end{array} \quad \Longrightarrow \quad \begin{array}{r} N_{11} = Y_1 \\ N_{12} = Y_2 - Y_1 \\ N_{21} = Y_3 - Y_1 \\ N_{22} = (Y_4 - Y_3) - (Y_2 - Y_1) \end{array}$$

Obsérvese que Y_2 es el total del renglón 1 ($N_{1\bullet}$), Y_3 es el total de la columna 1 ($N_{\bullet 1}$) y Y_4 es el gran total ($N_{\bullet\bullet}$). Sabemos que la función de densidad del vector $(N_{11}, N_{12}, N_{21}, N_{22})$ es

$$f_{N_{11},N_{12},N_{21},N_{22}}(n_{11}, n_{12}, n_{21}|n_{22}) = \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{e^{-\lambda_{ij}}\lambda_{ij}^{n_{ij}}}{n_{ij}!}$$

Que en términos del vector de $Y's$ queda como:

$$\frac{e^{-\lambda_{11}}\lambda_{11}^{y_1}}{y_1!}\,\frac{e^{-\lambda_{12}}\lambda_{12}^{y_2-y_1}}{(y_2-y_1)!}\,\frac{e^{-\lambda_{21}}\lambda_{21}^{y_3-y_1}}{(y_3-y_1)!}\,\frac{e^{-\lambda_{22}}\lambda_{22}^{(y_4-y_3)-(y_2-y_1)}}{((y_4-y_3)-(y_2-y_1))!}$$

Que simplificando queda como:

$$\frac{\left(\frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}}\right)^{y_1}\left(\frac{\lambda_{12}}{\lambda_{22}}\right)^{y_2}\left(\frac{\lambda_{21}}{\lambda_{22}}\right)^{y_3}{\lambda_{22}}^{y_4}}{y_1!\,(y_2-y_1)!\,(y_3-y_1)!\,((y_4-y_3)-(y_2-y_1))!}\,e^{\lambda_{\bullet\bullet}} \tag{2}$$

Observaciones: Bajo independencia de las distribuciones Poisson involucradas, tenemos:

$$\lambda_{ij} = \lambda_{i\bullet}\,\lambda_{\bullet j}, \quad y \quad \frac{\lambda_{11}\,\lambda_{22}}{\lambda_{12}\,\lambda_{21}} = 1 \quad {}^{1}$$

Además...

$$\begin{aligned} \lambda_{\bullet\bullet} = \lambda_{11} + \lambda_{12} + \lambda_{21} + \lambda_{22} &= \lambda_{1\bullet}\lambda_{\bullet 1} + \lambda_{1\bullet}\lambda_{\bullet 2} + \lambda_{2\bullet}\lambda_{\bullet 1} + \lambda_{2\bullet}\lambda_{\bullet 2} \\ &= (\lambda_{1\bullet} + \lambda_{2\bullet})(\lambda_{\bullet 1} + \lambda_{\bullet 2}) \end{aligned}$$

Entonces, la distribución marginal de (Y_2, Y_3, Y_4) se obtiene sumando en (1) sobre el rango de y_1. La suma requerida es

$$\sum_{y_1} \frac{1}{y_1!\,(y_2-y_1)!\,(y_3-y_1)!\,((y_4-y_3)-(y_2-y_1))!}$$

Ajustemos esta expresión a la que utilizamos en la hipergeométrica. Definamos entonces, $y_1 = y, y_2 = n_1, y_3 = kyy_4 = n$. Entonces, bajo esta nueva notación

$$\begin{aligned} \sum_{y_1} \frac{1}{y_1!\,(y_2-y_1)!\,(y_3-y_1)!\,((y_4-y_3)-(y_2-y_1))!} &= \frac{1}{n_1!\,(n-n_1)!}\sum_{y}\binom{n_1}{y}\binom{n-n_1}{k-y} \\ &= \frac{1}{n_1!\,(n-n_1)!}\binom{n}{k} \end{aligned}$$

[1] Esta fracción es el Cociente de Momios y, bajo independencia, su valor es 1

y, por lo tanto nos queda:

$$\begin{aligned} f_{Y_2Y_3Y_4}(y_2,y_3,y_4) &= \frac{1}{y_2!\,(y_4-y_2)!}\binom{y_4}{y_3}\binom{\lambda_{12}}{\lambda_{22}}^{y_2}\binom{\lambda_{21}}{\lambda_{22}}^{y_3}\lambda_{22}^{y_4}\,e^{-\lambda_{\bullet\bullet}} \\ &= \binom{y_4}{y_2}\binom{y_4}{y_3}\binom{\lambda_{12}}{\lambda_{22}}^{y_2}\binom{\lambda_{21}}{\lambda_{22}}^{y_3}\lambda_{22}^{y_4}\frac{e^{-\lambda_{\bullet\bullet}}}{y_4!} \end{aligned} \tag{3}$$

con las restricciones $0 \leq y_2, y_3 \leq y_4$. Por lo que la densidad condicional de Y_1, dado, Y_2, Y_3, Y_4 es

$$\begin{aligned} f_{Y_1|Y_2Y_3Y_4}(y_1|y_2,y_3,y_4) &= \frac{f_{Y_1,Y_2Y_3Y_4}(y_1,y_2,y_3,y_4)}{f_{Y_2Y_3Y_4}(y_2,y_3,y_4)} \\ &= \frac{1}{y_1!\,(y_2-y_1)!\,(y_3-y_1)!\,((y_4-y_3)-(y_2-y_1))!}\,\frac{y_2!\,(y_4-y_2)!}{\binom{y_4}{y_2}} \\ &= \frac{\binom{y_2}{y_1}\binom{y_4-y_2}{y_3-y_1}}{\binom{y_4}{y_2}} \qquad max(0,y_3-(y_4-y_2) \leq\ y_1\ \leq\ \text{mín}\,(y_2\,,\,y_3) \end{aligned}$$

Que es una distribución hipergeométrica. Cualquier otra entrada de la tabla tiene la misma distribución con los parámetros correspondientes.

Reto: Utilizando el resultado (2) del desarrollo anterior, encontrar la distribución de $f_{Y_2Y_3|Y_4}(y_2,y_3|y_4)$, la cual corresponde a la distribución de los totales marginales.

2.2. DISTRIBUCIONES DE PROBABILIDAD ASOCIADAS A UNA TABLA DE CONTINGENCIA

Consideraremos únicamente una tabla 2x2, porque, la generalización para tablas más grandes es inmediata. Para ilustrar estas distribuciones haremos uso de un conjunto de datos sobre la relación entre sexo y ser o no ser bebedor frecuente. La tabla es la siguiente...

	Bebedor Frecuente		
Sexo	**SÍ**	**NO**	**TOTAL**
Hombres	1′630	5′550	7′180
Mujeres	1′684	8′232	9′916
TOTAL	3′314	13′782	17′096

2.2.1. DISTRIBUCIÓN CONJUNTA:

En este caso tenemos dos variables categóricas, $\mathbf{X}$ (sexo) y $\mathbf{Y}$ (bebedor frecuente), con $\mathbf{I} = 2$ y $\mathbf{J} = 2$ categorías, respectivamente. Su distribución conjunta es la probabilidad de que un sujeto seleccionado aleatoriamente obtenga un valor en el renglón i y en la columna j de la tabla. Es decir

$$\pi_{ij} = \mathbb{P}(\mathbf{X} = i, \mathbf{Y} = j), \qquad i,j = 1, 2$$

Que genera la distribución conjunta dada por la tabla

	Bebedor Frecuente	
Sexo	**SÍ**	**NO**
Hombres	π_{11}	π_{12}
Mujeres	π_{21}	π_{22}

Donde la forma de estimar las probabilidades π_{ij} es:

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}, \qquad i,j=1,2$$

Con lo que generamos la tabla

	Bebedor Frecuente	
Sexo	**SÍ**	**NO**
Hombres	0.095	0.325
Mujeres	0.099	0.482

Algunas de las formas en las que podemos interpretar estas probabilidades para un caso particular ($\hat{\pi}_{22}$) es:

- La probabilidad **estimada** de que una mujer no sea bebedora frecuente es 48.2 %.
- La probabilidad de que una persona sea mujer **y/que** no sea bebedora frecuente es 48.2 %.
- De las personas encuestadas 48.2 % resultaron ser mujeres **y/que** no eran bebedoras frecuentes.
- De las 17′096 personas encuestadas 48.2 % resultaron ser mujeres **y/que** no eran bebedoras frecuentes ***(Para sacarse un 10 ;) jejeje)***

2.2.2. DISTRIBUCIÓN MARGINALES:

Las distribuciones marginales son aquellas que ignoran la presencia o desagregación de la otra variable presente, esto se logra al sumar por renglón o columna las probabilidades conjuntas. En concreto

$$\pi_{i\bullet} = \pi_{i1} + \pi_{i2} = \sum_j \pi_{ij} \qquad y \qquad \pi_{\bullet j} = \pi_{1j} + \pi_{2j} = \sum_i \pi_{ij} \qquad i,j=1,2$$

Con tabla de distribuciones marginales dada por

	Bebedor Frecuente		
Sexo	**SÍ**	**NO**	**TOTAL**
Hombres	π_{11}	π_{12}	$\pi_{1\bullet}$
Mujeres	π_{21}	π_{22}	$\pi_{2\bullet}$
TOTAL	$\pi_{\bullet 1}$	$\pi_{\bullet 2}$	1

Además tenemos que las sumas de todas las celdas en la distribución conjunta, es igual a la unidad, así como la suma de las probabilidades marginales de las filas o las columnas.

$$\sum_i \pi_{i\bullet} = \sum_j \pi_{\bullet j} = \sum_i \sum_j \pi_{ij} = 1$$

Con nuestros datos, esta tabla queda como

	Bebedor Frecuente		
Sexo	**SÍ**	**NO**	**TOTAL**
Hombres	0.095	0.325	**0.42**
Mujeres	0.099	0.482	**0.58**
TOTAL	**0.194**	**0.806**	**1**

La última de las distribuciones asociada a una tabla de contingencia son las distribuciones condicionales.

2.2.3. DISTRIBUCIONES CONDICIONALES

En la mayoría de las tablas de contingencia, una variables, digamos Y, es una variable de respuesta y la otra (X) es una variable explicativa. Cuando X es fijo en vez de aleatorio, la noción de distribución conjunta para X y Y no es más significativa. Sin embargo, **para una categoría fija X**, Y tiene una distribución de probabilidad. Lo que mide este concepto es qué tanto cambia esta distibución en la medida en que X cambia. Dado un sujeto clasificado en el renglón i de X, $\pi_{j|i}$ denota las probabilidad de clasificación en la columna j de Y, j=1,2,...J.

$$\pi_{j|i} = \mathbb{P}(\mathbf{Y} = j|\mathbf{X} = i), \qquad i, j = 1, 2$$

El conjunto de probabilidades $\left\{\pi_{1|i}\,,\, \pi_{2|i}\,, ...,\, \pi_{J|i}\right\}$ constituyen la distribución condicional de **Y** en cada categoría i de **X**. Con tabla de distribuciones condicionales

DISTRIBUCIONES CONDICIONALES (X,Y)

	Bebedor Frecuente		
Sexo	**SÍ**	**NO**	**TOTAL**
Hombres	$\pi_{1\|1}$	$\pi_{2\|1}$	1
Mujeres	$\pi_{1\|2}$	$\pi_{2\|1}$	1

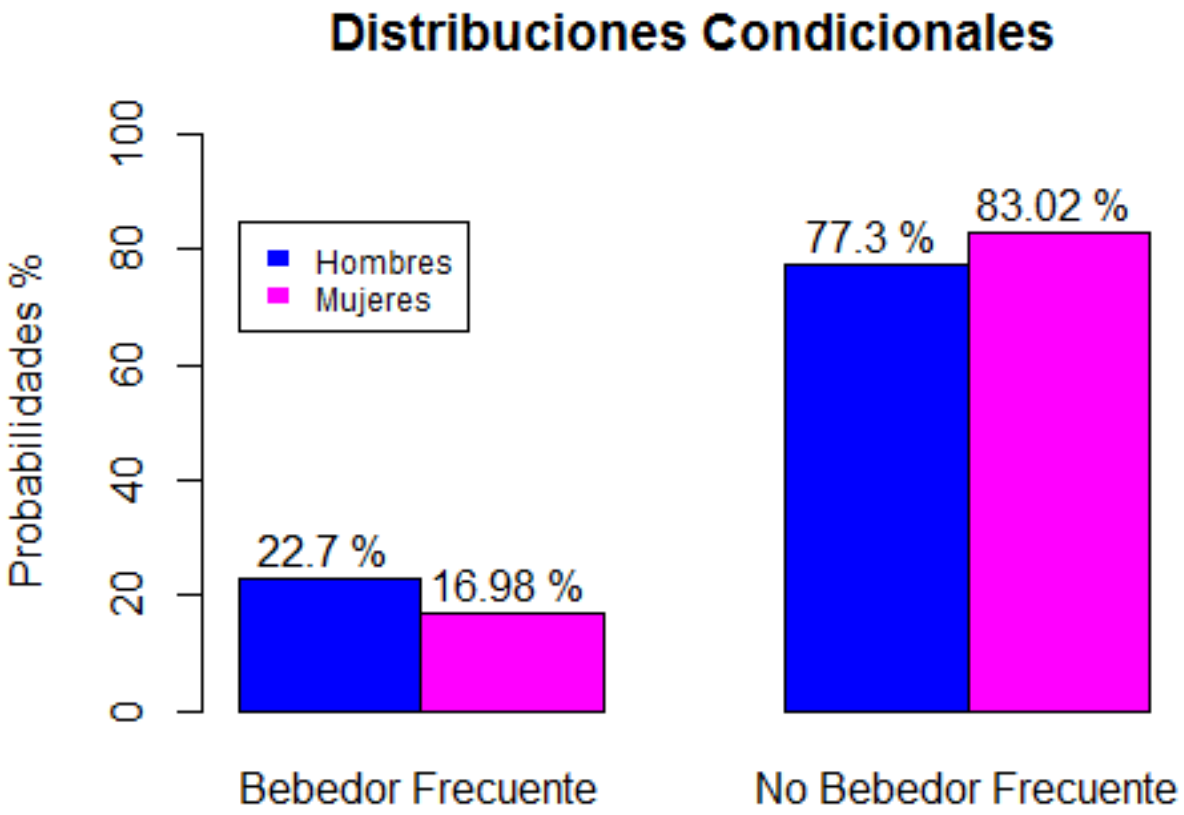
Las probabilidades estimadas en esta tabla están dadas por

$$\hat{\pi}_{j|i} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i\bullet}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i\bullet}}{n}} = \frac{n_{ij}}{n_{i\bullet}}, \qquad i, j = 1, 2$$

En nuestro caso, esta tabla queda

Distribuciones Condicionales (X,Y)

	Bebedor Frecuente		
Sexo	**SÍ**	**NO**	**TOTAL**
Hombres	0.2270195	0.7729805	1
Mujeres	0.1698265	0.8301735	1



Es interesante ver que los hombres (condicional hombre) tienen una probabilidad estimada mayor a ser bebedores frecuentes que las mujeres (condicional mujer). Observamos también que las distribuciones condicionales por sexo, no son muy diferentes. No obstante, aún no tenemos elementos para decidir de manera formal si son o no diferentes.

El objetivo principal de varios estudios es comparar las distribuciones condicionales de $\mathbf{Y}$ en varios niveles o categorías de variables explicativas.

En la siguiente tabla podemos resumir la estructura de las diferentes distribuciones de probabilidad de las tablas de contingencia

	Columnas		
Renglones	1	2	Total
1	π_{11}	π_{12}	$\pi_{1\bullet}$
	$\pi_{1\|1}$	$\pi_{2\|1}$	1.00
2	π_{21}	π_{22}	$\pi_{2\bullet}$
	$\pi_{1\|2}$	$\pi_{2\|2}$	1.00
Total	$\pi_{\bullet 1}$	$\pi_{\bullet 2}$	1.00

Tabla 2: Estructura General de las Distribuciones

2.2.4. Independencia de las Variables Categóricas

Una vez conociendo las distribuciones marginal, conjunta y condicional, estamos en posibilidades de ver la asociación entre ellas. Por ejemplo, las distribución condicional de Y dado X se relaciona con la distribución conjunta y marginal de la siguiente forma:

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i\bullet}} \qquad \forall i,j.$$

Dos variables de respuesta son definidas *independientes* si todas las probabilidades conjuntas igualan al producto de sus probabilidades marginales,

$$\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \qquad para \quad i = 1, 2, ...I \quad y \quad j = 1, 2, ...J$$

Cuando X y Y son independientes,

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i\bullet}} = \frac{(\cancel{\pi_{i\bullet}}\,\pi_{\bullet j})}{\cancel{\pi_{i\bullet}}} = \pi_{\bullet j} \qquad para \quad i = 1, ...I$$

Cada distribución condicional de Y es idéntica a la marginal de Y. Así que, dos variables son independientes cuando $\{\pi_{j|1} = \cdots = \pi_{j|I} \quad | \quad j = 1, ...J\}$; esto es, la probabilidad en cualquier columna será la misma para cada renglón. La Independencia suele ser entonces referida como una homogeneidad de las distribuciones condicionantes.

3. Medidas de Asociación en las Tablas de Contingencia

Ahora definiremos varias medidas de asociación relacionadas con estas tablas de contingencia que hemos presentado. La pregunta fundamental cuando tenemos dos variables categóricas es si éstas son independientes o no. En nuestro ejemplo, quisiéramos averiguar si el ser bebedor frecuente está relacionado con el sexo. Entonces, lo primero que debemos hacer es verificar si efectivamente estas variables están relacionadas.

3.1. Prueba Ji-Cuadrada (de Pearson) de Independencia

En 1900 el eminente estadístico Británico Karl Pearson introdujo una prueba de hipótesis que fue uno de los primeros métodos inferenciales. Éste tuvo un impacto revolucionario en el análisis de datos categóricos, que se enfocó en describir asociaciones. **La prueba de Pearson evalúa si los parámetros multinomiales igualan ciertos valores específicos.**

Para realizar esta prueba debemos definir, inicialmente, cuáles son las hipótesis que hay que contrastar para determinar si las variables que conforman nuestra tabla de contingencia son o no independientes.

Recordemos que dentro de la metodología estadística para realizar pruebas de hipótesis, es necesario determinar las llamadas hipótesis nula ($\mathbb{H}_0$) y la hipótesis alternativa ($\mathbb{H}_a$). En este caso de independencia, las enunciamos como

$\mathbb{H}_0$: Las variables son independientes V.S. $\mathbb{H}_a$: Las variables NO son independientes

Ahora debemos "traducir" estas hipótesis a elementos estadísticos para realizar la prueba.

De los cursos básicos de probabilidad, sabemos que si dos variables son independientes, su probabilidad conjunta es el producto de sus probabilidades marginales, como vimos anteriormente. En nuestra notación:

$$\pi_{ij} = \pi_{i\bullet}\,\pi_{\bullet j} \quad ó \quad \mathbb{P}(X=i, Y=j) = \mathbb{P}(X=i)P(Y=j) \qquad i,j=1,2$$

Por lo que las hipótesis pueden escribirse como

$$\begin{gathered}\mathbb{H}_0 : \mathbb{P}(X=i, Y=j) = \mathbb{P}(X=i)P(Y=j) \\ V.S. \\ \mathbb{H}_a : \mathbb{P}(X=i, Y=j) \neq \mathbb{P}(X=i)P(Y=j)\end{gathered}$$

La manera de probar la hipótesis de independencia es comparando las frecuencias observadas de la muestra, con las que se esperarían observar si efectivamente estas variables fueran independientes, es decir, comparar lo que realmente observamos con lo que esperamos observar si la hipótesis nula (independencia) es cierta. Esta comparación la realizaremos utilizando la famosa Ji-cuadrada de Pearson. Para esto, necesitamos definir primero los valores esperados de nuestra tabla.

De acuerdo a la definición de valor esperado, para obtener los valores esperados de cada una de las celdas de nuestra tabla de contingencia, debemos multiplicar el número de sujetos en la muestra por la probabilidad de caer en la celda correspondiente, suponiendo que las variables son independientes, es decir, bajo la hipótesis nula $\mathbb{H}_0$. En este caso los valores esperados son:

$$\mathbb{E}_{ij} = n\,\pi_{i\bullet}\,\pi_{\bullet j}$$

Que podemos estimar como:

$$\mathbb{E}_{ij} = n\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}, \qquad i,j = 1,2$$

Con estos valores podemos construir la prueba *Ji-cuadrada de independencia*, comparándolos contra los valores que observamos, n_{ij} , de la siguiente manera:

$$\mathcal{X}^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \overset{a}{\sim} \mathcal{X}^2_{(1)}$$

Presentamos los desarrollos para tablas 2x2, mismos que se pueden extender fácilmente para tablas generales de **IxJ**. La estadística Ji-cuadrada quedaría en este caso como:

$$\mathcal{X}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \overset{a}{\sim} \mathcal{X}^2_{((I-1)(J-1))}$$

Por la forma de la estadística, es obvio que si el valor de la misma es grande, entonces los valores esperados y los observados difieren mucho, lo que implicaría que la hipótesis nula es falsa. De lo contrario, estos valores son similares, lo que implicaría que la hipótesis nula es cierta y las variables son independientes.

Reglas de uso de la Ji-cuadrada

La distribución Ji-cuadrada es una distribución continua que, en el caso de la prueba Ji-cuadrada de Pearson, aproxima a un proceso discreto, por lo que hay que verificar ciertas reglas de uso para garantizar que esta aproximación sea adecuada en una situación particular.

Tablas 2x2

- Si n $<$ 20, utilizar la prueba exacta de Fisher.
- Si n$\geq$20, utilizar la Ji-cuadrada si los valores esperados son mayores o iguales que 5. $\mathbb{E}_{ij} \geq 5$.

Tablas IxJ

- Usar la Ji-cuadrada si a lo más el 20 % de las celdas tiene $\mathbb{E}_{ij} < 5$, pero ninguna de ellas tiene un valor esperado menor que 1.
- Cuando no se cumpla esta regla, pueden agruparse (colapsarse) categorías, siempre y cuando esto tenga sentido.

Una manera de mejorar la aproximación a la Ji-cuadrada en este tipo de tablas, es incluir la corrección por continuidad de Yates (1934). No obstante, ahora es posible, gracias al desarrollo de métodos computacionales más eficientes, calcular la distribución exacta de esta estadística cuando se tienen muestras pequeñas.

3.1.1. Corrección por Continuidad de Yates

Reescribamos la tabla 2x2 de la siguiente manera

Tablas (X,Y)			
	y_1	y_2	Total
x_1	a	b	a+b
x_2	c	d	c+d
Total	a+c	b+d	a+b+c+d=n

Yates (1934), argumenta que la distribución $\mathcal{X}_1^2$ proporciona únicamente una aproximación de las probabilidades asociadas a estos datos discretos, y, por lo tanto, los p-values basados en la estadística Ji-cuadrada generalmente subestiman los verdaderos p-values. En este contexto, Yates sugiere que la Ji-cuadrada debe ser corregida por continuidad y propone la siguiente corrección

$$\mathcal{X}^2 = \frac{n\ (|ad - bc| - 1/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Ahora veamos si existe asociación entre la condición de bebedor frecuente y el sexo en nuestros datos. Entonces, nuestra tabla es

		Bebedor Frecuente		
		SÍ	NO	Total
	Hombres	1630	5550	7180
Sexo	Mujeres	1684	8232	9916
	Total	3314	13782	17096

y queremos probar la hipótesis

$\mathbb{H}_0$: La condición de bebedor frecuente y el sexo son independientes.

vs

$\mathbb{H}_a$: La condición de bebedor frecuente está asociada con el sexo.

Como mencionamos, es necesario construir los valores esperados para aplicar la prueba Ji-cuadrada. Recordemos que, de manera general, se calculan como $\mathbb{E}_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$; $\quad i, j = 1, 2$. Mostremos estos cálculos para nuestro ejemplo

$$\mathbb{E}_{11} = \frac{7180 * 3314}{17096} = 1391.818 \qquad \mathbb{E}_{12} = \frac{7180 * 13782}{17096} = 5788.182$$
$$\mathbb{E}_{21} = \frac{9916 * 3314}{17096} = 1922.182 \qquad \mathbb{E}_{22} = \frac{9916 * 13782}{17096} = 7993.818$$

que generan la tabla de valores esperados

Valores Esperados

		Bebedor Frecuente		
		SÍ	NO	Total
	Hombres	1391.818	5788.182	7180
Sexo	Mujeres	1922.182	7993.818	9916
	Total	3314	13782	17096

Podemos apreciar que se cumplen de sobra, las condiciones para aplicar la prueba Ji-cuadrada de independencia. Entonces, el cálculo de la misma es

$$\mathcal{X}^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \mathbb{E}_{ij})^2}{\mathbb{E}_{ij}} = \quad \frac{(1630 - 1391.818)^2}{1391.818} + \frac{(5550 - 5788.182)^2}{1391.818} + \frac{(1684 - 1922.182)^2}{1391.818} + \frac{(8232 - 7993.818)^2}{1391.818} = 87.1718$$

que proporciona un p-value: $\mathbb{P}(\mathcal{X}^2_{(1)} \geq 87.1718) \approx 0$. Por lo que concluimos que existe asociación entre ser bebedor frecuente y el sexo.

3.1.2. Efecto del tamaño de muestra sobre la Ji-Cuadrada

Consideremos las siguientes tablas extraídas de *Statistical Methods for the Social Sciences*, página 268. Estos datos muestran el resultado por raza, sobre una pregunta acerca de la legalización del aborto. En cada tabla, 49 % de individuos de raza blanca y 51 % de raza negra, se pronuncian a favor de legalizar esta práctica.

Tabla: A

	SÍ	NO	Total
Blanca	49	51	100
Negra	51	49	100
Total	100	100	200
$\mathcal{X}^2 = 0.08$			$p = 0.78$

Tabla: B

	SÍ	NO	Total
Blanca	98	102	200
Negra	102	98	200
Total	200	200	400
$\mathcal{X}^2 = 0.08$			$p = 0.689$

Tabla: C

	SÍ	NO	Total
Blanca	4900	5100	10000
Negra	5100	4900	10000
Total	10000	10000	20000
$\mathcal{X}^2 = 8$			$p = 0.0046$

Obsérvese como el valor de la Ji-cuadrada va incrementándose de un valor que es altamente no significativo (p=0.78) hasta uno que es muy significativo (p=0.0046), conforme el tamaño de muestra crece de 200 a 400 y 20000. Ya que el p-value para esta estadística de asociación es muy sensible al tamaño de muestra, necesitamos otras medidas que describan la fuerza de esta asociación.

3.2. Prueba exacta de Fisher

Cuando en una tabla 2x2 no se cumplen las condiciones para utilizar la prueba Ji-cuadrada de independencia, tamaño de muestra pequeño o ($\mathbb{E}_{ij} < 5$), se puede utilizar la llamada prueba exacta de Fisher. La prueba se basa en el hecho de que la probabilidad exacta de observar una tabla con celdas: a, b, c y d corresponde a una distribución hipergométrica. Es importante remarcar que esta prueba asume que los marginales son fijos, lo que permite encontrar la distribución de la tabla únicamente a través de una de sus celdas. Nuevamente consideremos la tabla

Tablas (X,Y)			
	y_1	y_2	Total
x_1	a	b	a+b
x_2	c	d	c+d
Total	a+c	b+d	a+b+c+d=n

Entonces

$$\mathbb{P}(a,b,c,d) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Ejemplo: La catadora de té

Ilustraremos el proceso de cómo funciona esta prueba exacta, con el mismo experimento y datos coleccionados por el propio Fisher. El experimento, conocido como la catadora de té, consiste en averiguar si una persona puede determinar si en una taza de té se vertió primero la leche y después el té o viceversa. Entonces, se le proporcionó al azar a una catadora ocho tazas de té, de las cuales en cuatro de ellas se puso primero la leche y en las otras cuatro primero el té. La tabla asociada a este experimento es

Catadora de té

		Opinión		
		Leche	Té	Total
	Leche	3	1	4
Primero	Té	1	3	4
	Total	4	4	8

En este caso, las hipótesis a contrastar son

$\mathbb{H}_0$: No hay asociación entre la opinión de ella y lo que realmente ocurrió: Sólo adivina.

vs

$\mathbb{H}_a$: Si hay relación: Ella realmente posee esta habilidad para discernir

Entonces, dados los marginales fijos, la probabilidad asociada a estos datos observados se puede encontrar calculando, por medio de una hipergeométrica, la probabilidad asociada a la primera entrada. En concreto

$$\mathbb{P}(a=3)=\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}}=0.22857$$

Para determinar si existe o no asociación entre la opinión de la catadora y lo que realmente ocurrió, o si se puede afirmar que ella sólo adivinaba, es necesario encontrar el p-value asociado.El proceso estándar para calcular este p-value es, primeramente, considerar el conjunto de todas las tablas que tienen exactamente los mismos marginales que la tabla observada

Catadora de té: tablas con marginales iguales

		Opinión		
		Leche	Té	Total
	Leche	3		4
Primero	Té			4
	Total	4	4	8

para, posteriormente, calcular este valor como

p-value $=\sum$ Probabilidades de tablas a favor de $\mathbb{H}_a$, incluyendo a la de los datos observados. Entonces, observemos que el único valor más extremo en la tabla observada, es cuando la primera celda sea a = 4. En este caso, la probabilidad asociada a este valor es

$$\mathbb{P}(a=0)=\frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}}=0.01429$$

De donde el p-value es: 0.22857+0.01429=0.24286, y concluimos que esta mujer sólo está adivinando y no posee ninguna habilidad especial para discernir si se vierte primero la leche o el té. Como dato informativo, el p-value asociado por medio de la Ji-cuadrada es: 0.4795, cercano al doble de la prueba exacta.

Observación: No siempre es fácil determinar cuáles son los valores de la primera celda que hacen más extrema esta tabla. A este respecto, es más fácil trabajar con el cociente de momios, θ (que presentaremos más adelante), ya que a través de él es más fácil determinar esta situación. Una vez que estimemos este cociente de momios con la tabla observada, los valores del p-value se calcularán como

- Si $\mathbb{H}_a$ es una prueba de cola izquierda ($\theta < 1$)
  p-value $=\sum$Probabilidades de tablas en donde ocurra que $\theta < 1$

- Si $\mathbb{H}_a$ es una prueba de cola derecha ($\theta > 1$)
  p-value $=\sum$Probabilidades de tablas en donde ocurra que $\theta > 1$

Medidas de Asociación

Cuando utilizamos la estadística Ji-cuadrada para probar independencia entre dos variables categóricas, nuestra conclusión es si éstas son o no independientes. En caso de que fueran independientes, afirmamos que no están asociadas o que su grado de asociación es nulo, pero cuando no son independientes, la Ji-cuadrada no proporciona ninguna medida para determinar este grado o fuerza de la asociación, ni tampoco la dirección de la misma. Esto hace necesario la introducción de medidas relacionadas a estas tablas de contingencia, que proporcionen la magnitud de esta asociación y su dirección.

Una forma de medir la asociación entre distintas categorías de estas tablas, es a través de los residuos asociados a la $\mathcal{X}^2$. Ya que esta estadística compara valores esperados bajo el supuesto de independencia, con los valores observados, entonces, la diferencia

$$n_{ij} - \mathbb{E}_{ij}$$

debería ser una medida de independencia en caso de se cercana a cero esta diferencia o una medida de la falta de independencia o , es decir, una medida de correlación o de asociación. Los residuos estandarizados y ajustados de estas tablas, son

$$e_{ij} = \frac{n_{ij} - \mathbb{E}_{ij}}{\sqrt{\mathbb{E}_{ij}}} \qquad z_{ij} = \frac{n_{ij} - \mathbb{E}_{ij}}{\sqrt{\mathbb{E}_{ij}\left(1 - \frac{n_{i\bullet}}{n}\right)\left(1 - \frac{n_{\bullet j}}{n}\right)}}$$

Obsérvese que los residuos estandarizados son, de hecho, la raíz cuadrada de cada uno de los elementos de la Ji-cuadrada, y los residuos ajustados, son una modificación de los estandarizados. Esta modificación tiene el objetivo de dar importancia a los elementos de la tabla que pertenecen a renglones o columnas con baja frecuencia. Entonces, su objetivo es que estas diferencias no se vean atenuadas por frecuencias bajas. Además, los residuos ajustados tienen una distribución aproximadamente normal para muestras grandes, lo que permite determinar cuáles de ellos pueden considerarse grandes. Con nuestros datos, estas tablas son

Residuos Estandarizados

		Bebedor Frecuente	
		SÍ	NO
Sexo	Hombres	6.38	-3.13
	Mujeres	5.43	2.66

Residuos Ajustados

		Bebedor Frecuente	
		SÍ	NO
Sexo	Hombres	9.34	-9.34
	Mujeres	-9.34	9.34

Si observamos la tabla de residuos ajustados, podemos ver que son iguales y sólo difieren en signo. Además, como son normales, los cuatro son estadísticamente significativos al nivel de significancia $(0.05/2, N(0,1) = 1.96)$. Y ¿cómo se interpretan estos residuos?

En los casos con residuo positivo tenemos que el valor esperado bajo independencia, es menor que el valor observado, mientras que cuando el residuo es negativo, tenemos valores esperados mayores que los observados. En ambos casos hay una dependencia en cada una de las categorías; entonces, los hombres están asociados de manera positiva con la condición de bebedor frecuente y negativa con la de no serlo, mientras que las mujeres están asociadas de manera negativa con ser bebedoras frecuentes, pero de manera positiva (débil) con no serlo.

Nota: Revisar gráficas de asociación (mosaico)

3.3. Comparación de Proporciones

Varios estudios están diseñados para comparar grupos con variables de respuesta binarias. Así que Y tiene sólo dos categorías, tales como éxito y fracaso de resultados de un tratamiento médico. Con dos grupos, una tabla de contingencia de 2x2 muestra los resultados. Las filas son los grupos y las columnas son las categorías de Y.

3.3.1. Diferencia de proporciones

Comparar proporciones es una manera simple de juzgar la fuerza de asociación entre dos variables categóricas. Si en una tabla 2x2 consideramos una de las variables, $\mathbf{Y}$, como respuesta y la otra, $\mathbf{X}$, como explicativa, entonces podemos caracterizar esta tabla de la siguiente manera:

		Y		
		y_1	y_2	Total
X	x_1	π_1	$1-\pi_1$	1.00
	x_2	π_2	$1-\pi_2$	1.00

donde:

- $\pi_1 = \mathbb{P}[Y=1|X=1] = \pi_{1|1}$ $\quad y \quad$ $1-\pi_1 = \mathbb{P}[Y=2|X=1] = \pi_{2|1} = 1-\pi_{1|1}$
- $\pi_2 = \mathbb{P}[Y=1|X=2] = \pi_{1|2}$ $\quad y \quad$ $1-\pi_2 = \mathbb{P}[Y=2|X=2] = \pi_{2|2} = 1-\pi_{1|2}$

Entonces definimos la diferencia de proporciones como una medida de asociación en este tipo de tablas, dada por

$$\pi_1 - \pi_2 = (1-\pi_2) - (1-\pi_1)$$

Observemos que:

- $-1 \leq \pi_1 - \pi_2 \leq 1$
- Si $\mathbf{X} \perp \mathbf{Y} \Rightarrow \pi_1 - \pi_2 = 0$
- **Si** $\pi_1 - \pi_2 > 0$ **entonces la asociación es positiva** en el sentido de que es más probable observar la respuesta $\mathbf{Y}=1$, en el renglón 1 ($\mathbf{X}=1$), que observar esta misma respuesta en el renglón 2, ($X=2$).
- **Si** $\pi_1 - \pi_2 < 0$ **entonces la asociación es negativa** en el sentido de que es más probable observar la respuesta $\mathbf{Y}=1$, en el renglón 2 ($\mathbf{X}=2$), que observar esta misma respuesta en el renglón 1, ($\mathbf{X}=1$).

Es interesante observar que la diferencia de proporciones está acotada en el rango $[-1,1]$. **Una diferencia cercana a uno en valor absoluto, indica un alto grado de asociación**, mientras que **una diferencia cercana a cero representa muy baja asociación**. Notemos la similitud de esta medida con el coeficiente de correlación para datos continuos (Correlación de Pearson). En nuestro caso

π_1 : Proporción de hombres que son bebedores frecuentes
π_2 : Proporción de mujeres que son bebedoras frecuentes

Esta diferencia de proporciones estimada, es:

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{n_{11}}{n_{1\bullet}} - \frac{n_{21}}{n_{2\bullet}} = \frac{1630}{1630 + 5550} - \frac{1684}{1684 + 8232} = 0.227 - 0.170 = 0.057$$

Que sugiere un nivel de asociación muy bajo.

Finalmente, para determinar si la magnitud de esta diferencia en las proporciones puede considerarse estadísticamente distinta de cero, se realiza la prueba correspondiente.

Nota: Estudiar la triada de pruebas "Wald-Likelihood Ratio-Score"

Prueba de Hipótesis:

$$\mathbb{H}_0 : \pi_1 = \pi_2 = \pi \qquad vs \qquad \mathbb{H}_1 : \pi_1 \neq \pi_2$$

Bajo $\mathbb{H}_0$

$$\hat{\pi} = \frac{n_{1\bullet}\hat{\pi}_1 + n_{2\bullet}\hat{\pi}_2}{n_{1\bullet} + n_{2\bullet}}$$

y, bajo $\mathbb{H}_0$, sabemos que, para muestras grandes tenemos la distribución asintótica

$$\hat{\pi}_1 - \hat{\pi}_2 \approx N\left(\pi_1 - \pi_2, \frac{\hat{\pi}\,(1-\hat{\pi})}{n_{1\bullet}} + \frac{\hat{\pi}\,(1-\hat{\pi})}{n_{2\bullet}}\right)$$

Con lo que la estadística de prueba queda como

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_{1\bullet}} + \frac{1}{n_{2\bullet}}\right)}} \approx N(0,1) \quad \acute{o} \quad Z^2 \approx \mathcal{X}^2_{(1)}$$

Con nuestros datos

$$\hat{\pi} = \frac{7180 * 0.227 + 9916 * 0.17}{7180 + 9916} = 0.194$$

$$Z = \frac{0.227 - 0.17}{\sqrt{0.194(1-0.194)\left(\frac{1}{7180} + \frac{1}{9916}\right)}}$$

$$Z^2 = 9.302^2 = 86.52$$

y concluimos que la proporción de hombres que son bebedores frecuentes es diferente de la proporción de mujeres que lo son.

3.3.2. Corrección Binomial por Continuidad

Observemos que los datos involucrados en estas pruebas corresponden a distribuciones binomiales, y que nuevamente se asocia una distribución continua (normal o ji-cuadrada) a un proceso discreto (binomial). Por lo que podemos utilizar una corrección para esta estadística, dada por

$$Z_c = \frac{|\hat{\pi}_1 - \hat{\pi}_2| \left(\frac{1}{2\, n_{1\bullet}} + \frac{1}{2\, n_{2\bullet}} \right)}{\sqrt{\hat{\pi}(1-\hat{\pi}) \left(\frac{1}{n_{1\bullet}} + \frac{1}{n_{2\bullet}} \right)}} \approx N(0,1) \quad ó \quad Z_c^2 \approx \mathcal{X}^2_{(1)}$$

3.3.3. Estimación del Intervalo para la Diferencia de Proporciones

La diferencia de proporciones compara las distribuciones condicionales de las variables de respuesta para dos grupos. Para estas mediciones, tratamos a las muestras como binomiales independientes. Para el grupo i, y_i tiene una distribución binomial con tamaño de muestra $n_{i\bullet}$, y probabilidad π_i de éxitos. La proporción muestral $\hat{\pi}_i = y_i/n_{i\bullet}$ tiene como valor esperado π_i y varianza $\pi_i(1-\pi_i)/n_{i\bullet}$. Como $\hat{\pi}_1$ y $\hat{\pi}_2$ son independientes, su diferencie tiene

$$\mathbb{E}(\hat{\pi}_1 - \hat{\pi}_2) = \pi_1 - \pi_2$$

Y error estándar

$$\sigma(\hat{\pi}_1 - \hat{\pi}_2) = \left[\frac{\pi_1(1-\pi_1)}{n_{1\bullet}} + \frac{\pi_2(1-\pi_2)}{n_{2\bullet}} \right]^{1/2}.$$

La estimación $\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2)$ usa la fórmula anterior reemplazando π_i por $\hat{\pi}_i$. Entonces

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\, \hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2)$$

es el intervalo de confianza de Wald para $\pi_1 - \pi_2$. Al igual que el intervalo de confianza de Wald para una única proporción, éste usualmente tiene una probabilidad de cobertura real menor que el coeficiente nominal de confianza, especialmente cuando π_1 y π_2 son cercanos a 0 o 1.

Existen algunos otros métodos para estimar el intervalo de confianza de la diferencia de proporciones.

Para dos proporciones, un intervalo de confianza para $\pi_1 - \pi_2$ basado en la estimación muestral única $\hat{\pi}_i$ y el intervalo (l_i, u_i) para π_i, $i = 1, 2$, es

$$\left(\hat{\pi}_1 - \hat{\pi}_2 - \sqrt{(\hat{\pi}_1 - l_1)^2 + (u_2 - \hat{\pi}_2)^2}\ ,\ \hat{\pi}_1 - \hat{\pi}_2 + \sqrt{(u_1 - \hat{\pi}_1)^2 + (\hat{\pi}_2 - l_2)^2} \right).$$

En 1998 Newcombe propuse un intervalo para $\pi_1 - \pi_2$ usando el intervalo score (l_i, u_i) para π_i que funciona mucho mejor que el intervalo de Wald propuesto. Éste es

$$(\hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2} s_L,\ \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2} s_U), \qquad con$$

$$s_L = \sqrt{\frac{l_1(1-l_1)}{n_{1\bullet}} + \frac{u_2(1-u_2)}{n_{2\bullet}}} \quad , \quad s_U = \sqrt{\frac{u_1(1-u_1)}{n_{1\bullet}} + \frac{l_2(1-l_2)}{n_{2\bullet}}}.$$

3.4. Riesgo Relativo (R.R.)

Una diferencia de proporciones fija, puede ser más relevante cuando los valores de las probabilidades están cercanos a cero o uno, que cuando están en el rango medio (0.5). Por ejemplo, la diferencia entre 0.10 y 0.01 es 0.09, que es igual a la diferencia entre 0.5 y 0.41, sin embargo, en el primer caso una es 10 veces más grande que la otra, mientras que en el segundo caso es tan sólo de 1.2. En este caso, es mejor utilizar el llamado riesgo relativo, como una medida de asociación que muestra de mejor manera esta comparación entre riesgo (probabilidades) muy bajos. La definición del riesgo relativo entre sujetos expuestos vs. no expuestos[2], es

$$RR = \frac{\text{Probabilidad de desarrollar la enfermedad para individuos expuestos}}{\text{Probabilidad de desarrollar la enfermedad para individuos no expuestos}} = \frac{\pi_1}{\pi_2}$$

cuyo estimador es

$$\widehat{RR} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Entonces tenemos que

- **RR** > 0
- **RR** $\approx 1 \Rightarrow$ La asociación entre exposición y riesgo es probable que no exista, es decir, las variables no están correlacionadas o asociadas.
- **RR** $> 1 \Rightarrow$ El riesgo se incrementa entre aquellos sujetos expuestos.
- **RR** $< 1 \Rightarrow$ El riesgo decrece entre aquellos sujetos expuestos.

En nuestro caso, el riesgo relativo de ser bebedor frecuente entre los dos sexos es

$$RR = \frac{\frac{1630}{1630+5550}}{\frac{1684}{1684+8232}} = 1.337$$

Interpretación La estimación anterior se interpreta como: "*El riesgo de ser bebedor frecuente es* 33.7 % *más grande en los hombres que en las mujeres*".

Note que la fuerza de asociación no es muy grande entre estas variables. Para realizar inferencias necesitamos conocer la varianza de este estimador. Para ello haremos uso de un resultado sumamente útil llamado el *método delta*, que enunciaremos a continuación.

[2]En nuestro caso, el factor de exposición es el sexo, y el factor de riesgo es bebedor frecuente

3.4.1. Método Delta

Sea X una v.a. y sea g una función infinitamente diferenciable, esto es, g tiene derivadas de cualquier orden. Supongamos que el cálculo exacto de la esperanza y varianza de g(X), es muy complicado. El método delta se utiliza para aproximar esta esperanza y varianza, utilizando **series de Taylor**.

Recordatorio: Series de Taylor

Los polinomios figuran entre las funciones más sencillas que se estudian en análisis. Son adecuadas para trabajar en cálculos numéricos porque sus valores se pueden obtener efectuando un número finito de multiplicaciones y adiciones. Existen muchas funciones como la logarítmica, la exponencial, las trigonométricas, etc que pueden aproximarse por polinomios, lo que permite calcular las funciones originales con la precisión que se desee. Si la diferencia entre una función y su aproximación polinómica es suficientemente pequeña, entonces podemos, a efectos prácticos, calcular con el polinomio en lugar de hacerlo con la función original.

Podemos demostrar que existe un polinomio y uno sólo de grado $\leq n$ que coincide con f y sus primeras n derivadas en el punto $x = a$. Podemos entonces escribir P(x) ordenado según las potencias de $x - a$ y proceder a hacer la aproximación. Si calculamos las derivadas en el punto "a", llegamos al polinomio

$$P(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

y se le llama ***polinomio de Taylor de grado n generado por f en el punto a***.

Continuando con el método Delta... Denotemos $\mathbb{E}(X) = \mu$ y $\mathbb{V}(X) = \sigma^2$. Entonces, para aproximar el valor de la esperanza de g(X), utilizando la expansión de g(X) alrededor de μ, tenemos

$$\begin{aligned} g(X) &\approx g(\mu) + (X-\mu)\left.\frac{dg}{dX}\right|_{X=\mu} \\ &= g(\mu) + (X-\mu)\, g'(\mu) \end{aligned}$$

tomando esperanzas, tenemos

$$\begin{aligned} \mathbb{E}[\, g(X)\,] &\approx \mathbb{E}[\, g(\mu) + (X-\mu)\, g'(\mu)\,] \\ &= \mathbb{E}[\, g(\mu)\,] + \mathbb{E}[\,(X-\mu) g'(\mu)\,] \\ &= g(\mu) + g'(\mu)\mathbb{E}[\, X-\mu\,] \\ &= g(\mu) + g'(\mu)\cdot 0 \\ &= g(\mu) \end{aligned}$$

Para obtener la aproximación de su varianza, nuevamente desarrollamos alrededor de μ la expansión de Taylor, y tenemos

$$g(X) \approx g(\mu) + (X-\mu)\, g'(\mu), \text{ entonces,}$$

$$\begin{aligned}
\mathbb{V}(g(X)) &= \mathbb{E}[\,(g(X)-\mathbb{E}(\mathbb{X}))^2\,] \\
&\approx \mathbb{E}[\,(g(X)-g(\mu))^2\,] \\
&\approx \mathbb{E}[\,(\cancel{g(\mu)}+(X-\mu)\,g'(\mu)-\cancel{g(\mu)})^2\,] \\
&= \mathbb{E}[\,((X-\mu)\,g'(\mu))^2\,] \\
&= g'(\mu)^2\,\mathbb{E}[\,(X-\mu)^2\,] \\
&= g'(\mu)^2\,\sigma^2
\end{aligned}$$

Estas aproximaciones serán fundamentales para encontrar la varianza del RR. Antes de comenzar el desarrollo de esta varianza, notemos lo siguiente

- Además de calcular la varianza del riesgo relativo, necesitaremos conocer su distribución, al menos, asintótica.

- En general, el logaritmo de una variable aleatoria tiene distribución más cercana a una normal, que la propia variable.

- Nuestro desarrollo se realizará sobre el logaritmo del RR y no sobre éste.

3.4.2. Calculo de la Varianza del Riesgo Relativo

$$\begin{aligned}
\mathbb{V}\left[\,log\left(\widehat{RR}\right)\,\right] &= \mathbb{V}\left[\,log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)\,\right] \\
&= \mathbb{V}[\,log(\hat{\pi}_1)-log(\hat{\pi}_2)\,] \\
&= \mathbb{V}[\,log(\hat{\pi}_1)\,]+\mathbb{V}[\,log(\hat{\pi}_2)\,]
\end{aligned}$$

Para encontrar las varianzas de los logaritmos de los estimadores, utilizaremos el método delta, con $g(X)=log(X)$, por lo que, $g'(X)=\frac{1}{X}$. Con esto obtenemos

$$\mathbb{V}(\hat{\pi}_i)=\frac{1}{\hat{\pi}_i^2}\,\mathbb{V}(\hat{\pi}_i)=\frac{1}{\hat{\pi}_i^2}\left(\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_{i\bullet}}\right)=\frac{1-\hat{\pi}_i}{n_{i\bullet}\,\hat{\pi}_i},\quad i=1,2$$

Dados los elementos de la tabla $n_{1\bullet}=a+b,\quad n_{2\bullet}=c+d,\quad \hat{\pi}_1=\dfrac{a}{a+b}$ y $\hat{\pi}_2=\dfrac{c}{c+d}$. Entonces

$$\mathbb{V}(log(\hat{\pi}_1))=\frac{1-\dfrac{a}{a+b}}{(a+b)\,\dfrac{a}{a+b}}=\frac{1-\dfrac{a}{a+b}}{a}=\frac{1}{a}-\frac{1}{a+b}$$

y

$$\mathbb{V}(log(\hat{\pi}_2))=\frac{1-\dfrac{c}{c+d}}{(c+d)\,\dfrac{c}{c+d}}=\frac{1-\dfrac{c}{c+d}}{c}=\frac{1}{c}-\frac{1}{c+d}$$

por lo que

$$\mathbb{V}\left(\,log(\widehat{RR})\right)=\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}$$

3.4.3. Intervalo de Confianza para el Riesgo Relativo

Si queremos un intervalo de confianza para el riesgo relativo, partimos del hecho de que

$$log(\widehat{RR}) \approx N\left(log(RR), \mathbb{V}\left[log\left(\widehat{RR}\right)\right]\right), \qquad n \longrightarrow \infty$$

y el intervalo de confianza Wald para el logaritmo del riesgo relativo queda como

$$log\left(RR\right) \in \left(log\left(\widehat{RR}\right) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right)$$

Tomando en cuenta que la función exponencial es estrictamente creciente, el intervalo de confianza para el riesgo relativo es

$$RR \in \left(exp\left\{log\left(\widehat{RR}\right) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}\right\}\right)$$

Para calcular el intervalo de confianza de nuestro RR, primero hay que calcular la varianza de su logaritmo, que es

$$\mathbb{V}\left(log(\widehat{RR})\right) = \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d} = \frac{1}{1630} - \frac{1}{7180} + \frac{1}{1684} - \frac{1}{9916} = 0.0009671983, \quad y$$

$$\mathbb{E}.\mathbb{S}.\left(log(\widehat{RR})\right) = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}} = \sqrt{\frac{1}{1630} - \frac{1}{7180} + \frac{1}{1684} - \frac{1}{9916}} = 0.03109981$$

Por lo que el intervalo de confianza al 95 % para este riesgo relativo es

$$RR \in exp\left\{\, log(1.337) \pm 1.96 * 0.03109981 \,\right\} = (1.2577, 1.4208)$$

3.5. Momio (odd)

Si algún evento ocurre con probabilidad p, entonces el *momio a favor de este evento* es

$$O = \frac{p}{1-p}$$

Por ejemplo, si $p = 1/2$ este momio será uno, e indica iguales oportunidades de que ocurra el evento a que no ocurra. Si $p = 2/3$, entonces el momio es

$$O = \frac{\frac{2}{3}}{1-\frac{2}{3}} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2$$

e implica el doble de posibilidades de que ocurra el evento a que no ocurra. En concreto, **un momio es la comparación de las probabilidades de observar una respuesta y no observar esta respuesta, entre individuos con la misma característica**. En nuestro caso, la probabilidad de observar la respuesta, $Y = 1$, contra la probabilidad de no observarla, $Y = 2$, en el renglón $X = 1$ ó $X = 2$. En símbolos

$$\Omega_1 = \frac{\mathbb{P}[Y=1|X=1]}{\mathbb{P}[Y=2|X=1]} = \frac{\pi_1}{1-\pi_1} \qquad \Omega_2 = \frac{\mathbb{P}[Y=1|X=2]}{\mathbb{P}[Y=2|X=2]} = \frac{\pi_2}{1-\pi_2}$$

Características de los momios

- $\Omega_i > 0$
- Si $\Omega_1 = \Omega_2$, entonces las variables son independientes
- Si Ω_1 es mayor que 1, entonces $\mathbb{P}[Y=1|X=1] > \mathbb{P}[Y=2|X=1]$, y hay una asociación positiva con la respuesta 1 (Y=1).
- Si Ω_2 es mayor que 1, entonces $\mathbb{P}[Y=1|X=2] > \mathbb{P}[Y=2|X=2]$, y hay una asociación positiva con la respuesta 2 (Y=2).

En nuestro ejemplo los momios corresponden a la "respuesta" de bebedor frecuente entre los hombres y entre las mujeres. Que calculamos como

$$\Omega_1 = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b} = \frac{1630}{5550} = 0.294 \qquad \Omega_2 = \frac{\frac{c}{c+d}}{\frac{d}{c+d}} = \frac{c}{d} = \frac{1684}{8232} = 0.205$$

Para interpretar mejor estos riesgos relativos, es más conveniente hacerlo con sus recíprocos $\frac{1}{\Omega_1}$ y $\frac{1}{\Omega_2}$, que son respectivamente: 3.40 y 4.87. Entonces, podemos decir que el riesgo relativo de no ser bebedor frecuente entre los hombres es de 3.4 respecto a ser bebedor frecuente, mientras que para las mujeres este mismo riesgo es 4.87. Podemos afirmar que hay una asociación más fuerte con la respuesta de no ser bebedor frecuente, que con la de serlo.

3.6. Cociente de Momios (Odds Ratio)

Una de las medidas de asociación de mayor uso es el llamado cociente de momios que, en este caso, es el cociente del momio de bebedor frecuente entre los hombres y el momio de bebedor frecuente entre las mujeres.

$$\theta = \frac{\Omega_1}{\Omega_2}$$

cuyas características son:

- *Independencia:* Cuando $\pi_1 = \pi_2$, entonces

$$\Omega_1 = \Omega_2 \Rightarrow \theta = \frac{\Omega_1}{\Omega_2} = 1$$

- *Dependencia:* Cuando $\pi_1 > \pi_2$, entonces

$$\Omega_1 > \Omega_2 \Rightarrow 1 < \theta < \infty$$

- *Dependencia:* Cuando $\pi_1 < \pi_2$, entonces

$$\Omega_1 < \Omega_2 \Rightarrow 0 < \theta < 1$$

La estimación del cociente de momios es

$$\hat{\theta} = \frac{\hat{\Omega}_1}{\hat{\Omega}_2} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

En nuestro caso

$$\hat{\theta} = \frac{1630 * 8232}{5550 * 1684} = 1.436$$

que se interpreta como el momio de bebedor frecuente en los hombres es 1.436 veces mayor que el momio de bebedora frecuente de las mujeres.

Si queremos realizar inferencias sobre el OR, debemos calcular su varianza y hacer uso de su distribución asintótica.

3.6.1. Cálculo de la Varianza del Cociente de Momios

Utilizaremos argumentos similares a los anteriores, para desarrollar la varianza del cociente de momios. En este caso, también trabajaremos con el logaritmo del **OR**. Para simplificar notación, usaremos, como es común, $\hat{q}_i = 1 - \hat{p}_i,\ \hat{p}_i = \hat{\pi}_i\ i = 1, 2.$

$$\begin{aligned}
\mathbb{V}\left(log\left(\widehat{OR}\right)\right) &= \mathbb{V}\left[log\left(\frac{\frac{\hat{p}_1}{\hat{q}_1}}{\frac{\hat{p}_2}{\hat{q}_2}}\right)\right] = \mathbb{V}\left[log\left(\frac{\hat{p}_1}{\hat{q}_1}\right) - log\left(\frac{\hat{p}_2}{\hat{q}_2}\right)\right] \\
&= \mathbb{V}\left[log\left(\frac{\hat{p}_1}{\hat{q}_1}\right)\right] + V\left[log\left(\frac{\hat{p}_2}{\hat{q}_2}\right)\right]
\end{aligned}$$

Para obtener la varianza de los términos involucrados en esta última expresión, utilizaremos el método delta. Observemos que el primer término sólo es función de $\hat{\pi}_1$ y el segundo de $\hat{\pi}_2$. Recordemos además que por el método Delta, $\mathbb{V}(g(X)) \approx \sigma^2\, g'(\mu)^2$. Por lo que tenemos

$$\frac{d\, log(\hat{p}_i/\hat{q}_i)}{d\,\hat{p}_i} = \frac{1}{\hat{p}_i/\hat{q}_i}\,\frac{d\,\hat{p}_i/\hat{q}_i}{d\,\hat{p}_i} = \frac{1}{\hat{p}_i\,\hat{q}_i} \quad i = 1, 2$$

Además, ya vimos anteriormente que

$$\mathbb{V}\,(\hat{p}_i) = \frac{\hat{p}_i\hat{q}_i}{n_{i\bullet}}, \quad i = 1, 2$$

por lo que

$$\begin{aligned} \mathbb{V}\left[log\left(\frac{\hat{p}_1}{\hat{q}_1}\right)\right] &\approx \frac{\hat{p}_1\hat{q}_1}{n_{1\bullet}}\left(\frac{1}{\hat{p}_1\hat{q}_1}\right)^2 = \frac{1}{\hat{p}_1\hat{q}_1 n_{1\bullet}} \\ &= \frac{1}{(a+b)\left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)} = \frac{a+b}{a\,b} = \frac{1}{a} + \frac{1}{b} \end{aligned}$$

Un desarrollo similar muestra que

$$\mathbb{V}\left[log\left(\frac{\hat{p}_2}{\hat{q}_2}\right)\right] \approx \frac{1}{c} + \frac{1}{d}$$

y, finalmente

$$\mathbb{V}\left(\widehat{OR}\right) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

3.6.2. Intervalo de Confianza para OR

Por teoría asintótica, tenemos que

$$log\left(\widehat{OR}\right) \approx N\left(log(OR), \mathbb{V}\left[log\left(\widehat{OR}\right)\right]\right), \quad n \to \infty$$

y el intervalo de confianza asintótico para el logaritmo del OR queda como

$$log(OR) \in \left(log\left(\widehat{OR}\right) \pm Z_{(1-\alpha/2)}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

el cuál es el intervalo de confianza Wald para $log\theta$. Entonces, el IC Wald para OR será

$$OR \in \left(exp\left\{log\left(\widehat{OR}\right) \pm Z_{(1-\alpha/2)}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right\}\right)$$

Calculemos la varianza para el cociente de momios asociado a nuestra tabla

$$\mathbb{V}\left(log\left(\widehat{OR}\right)\right) \approx \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d} = \frac{1}{1630}+\frac{1}{5550}+\frac{1}{1684}+\frac{1}{8232} = 0.001508979 \quad y,$$

$$\mathbb{E.S.}\left(log\left(\widehat{OR}\right)\right) \approx \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{1630}+\frac{1}{5550}+\frac{1}{1684}+\frac{1}{8232}} = 0.03884557$$

Con lo que el intervalo de confianza es

$$OR \in \{exp(log(1.436) \pm 1.96 * 0.03884557)\} = (1.330727, 1.549601)$$

En una tabla de 2x2, $\hat{\theta}$ será igual a cero o ∞ si cualquier $n_{ij} = 0$, y será indefinido si ambas entradas en un renglón o una columna son cero. Como estos resultados tienen probabilidades positivas, el valor esperado de $\hat{\theta}$ y $\log(\hat{\theta})$ no existirían. En términos de sesgo y error cuadrático medio, Gart y Zweiful (1967) y Haldane (1956) mostraron estimadores que enmiendan esto

$$\tilde{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{21}+0.5)(n_{12}+0.5)}$$

y así, $log(\tilde{\theta})$ tiene buen comportamiento.

Los estimadores $\hat{\theta}$ y $\tilde{\theta}$ tienen la misma distribución normal asintótica alrededor de θ. A no ser que n sea muy grande, sus distribuciones serán altamente sesgadas. Cuando $\theta = 1$, por ejemplo, $\hat{\theta}$ no puede ser mucho más pequeño que θ.

El intervalo de confianza de Wald construido a través del método Delta visto anteriormente fue propuesto por Woolf (1955). Cuando $\hat{\theta} = 0$ *o* ∞, este intervalo no existe. Cuando $\hat{\theta} = 0$, uno debería tomar 0 como el límite inferior y cuando $\hat{\theta} = \infty$, uno debería tomar ∞ como el límite superior. Otro intervalo propuesto puede usar la fórmula de Woolf siguiendo un ajuste, tal como el de Gart (1966), quien reemplaza $\{n_{ij}\}$ por $\{n_{ij}+0.05\}$ en el estimador y en el error estándar. Una forma menos ad hoc forma el intervalo invirtiendo la prueba score o la prueba del cociente de verosimilitudes para θ como vimos anteriormente.

Podemos observar que existen diversas medidas para analizar la asociación entre variables categóricas, y no se limitan únicamente al uso de la Ji-cuadrada de Pearson. Con las medidas que hemos introducido, se puede decidir si existe asociación entre estas variables y en qué sentido se da, pero resta tener una medida de la fuerza de esta asociación.