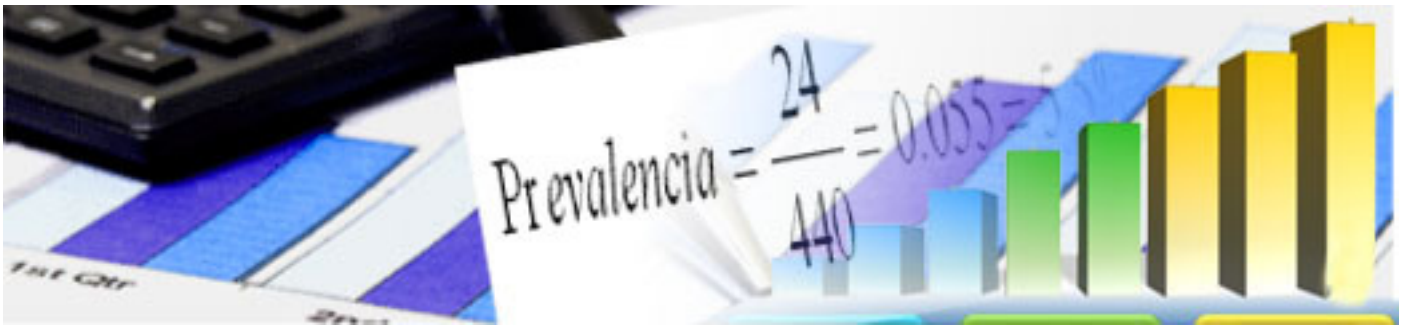
4.2 Conceptos básicos de Estadística II



Métodos para el estudio de la Proporción de individuos que presentan una determinada característica de interés en la población a estudio, a partir de la información proporcionada por los datos de la muestra.

Autora: Inma Jarrin Vera

Coordinadora de Estadística del Máster en Salud Pública

Se recomienda imprimir 2 páginas por hoja

Citación recomendada:

Jarrín Vera I. Conceptos básicos de Estadistica II [Internet]. Madrid: Escuela Nacional de Sanidad; 2012 [consultado día mes año]. Tema 4.2. Disponible en: direccion url del pdf.

Resumen:

En este tema se presentan los métodos analíticos utilizados para el análisis de variables de interés dicotómicas.

En primer lugar, se presentan dos tipos de proporciones que son de particular interés en la investigación médica: la prevalencia y la incidencia acumulada.

En segundo lugar, se describe cómo calcular la proporción de

individuos que presentan la característica de interés en los individuos de la muestra. Esta proporción se utiliza para determinar la proporción de individuos que presentan la característica de interés en la población a estudio. Para ello, se presenta la distribución en el muestreo de una proporción, conocida como distribución Binomial, y se muestra cómo esta distribución se aproxima a la distribución Normal, para el cálculo de un Intervalo de Confianza al 95% de la Proporción en la población.

En tercer lugar, se presentan los métodos para comparar la proporción de individuos que presentan la característica de interés en 2 grupos diferentes de individuos.

Y, por último, se extienden estos métodos para el caso en el que queremos comparar la proporción de individuos que presentan la característica de interés en más de 2 grupos diferentes de individuos.

Introducción

En la Investigación en Salud, son numerosas las ocasiones en las que nos planteamos estudiar variables de interés dicotómicas, es decir, variables categóricas que toman dos posibles valores. Por ejemplo, una mujer puede estar o no infectada por el virus

En la Investigación en Salud, son numerosas las ocasiones en las que nos planteamos estudiar variables de interés dicotómicas. En estas situaciones, resulta de particular interés determinar la proporción de individuos que presentan la característica de interés.

del papiloma humano; o un individuo infectado por VIH puede desarrollar SIDA o no a los 5 años de la seroconversión al VIH. En estas situaciones, resulta de particular interés determinar la proporción de individuos que presentan la característica de interés; por ejemplo, la proporción de mujeres de población general infectadas por el virus del papiloma humano.

Hay dos tipos de proporciones que son de particular relevancia en la Investigación Médica: ***Prevalencia*** *e* ***Incidencia Acumulada (Riesgo)***

1. Prevalencia e Incidencia Acumulada

Hay dos tipos de proporciones que son de particular relevancia en la Investigación Médica: Prevalencia e Incidencia Acumulada (Riesgo).

1.1. Prevalencia

La prevalencia es la carga de enfermedad en un momento puntual. Se calcula mediante la siguiente fórmula:

$$\text{Prevalencia} = \frac{\text{Número individuos con la enfermedad en un momento puntual}}{\text{Población total}}$$

La ***prevalencia*** *es la carga de enfermedad en un momento puntual*

Supongamos que estamos interesados en determinar la prevalencia de infección por VIH en las mujeres dedicadas a la prostitución en la Comunidad de Madrid. Para ello, seleccionamos una muestra de 440 mujeres, de las que 24 están infectadas por VIH. La Prevalencia de infección por VIH en las 440 mujeres de la muestra se calcularía como:

$$\text{Prevalencia} = \frac{24}{440} = 0.055 = 5.5\%$$

Entre las 440 mujeres de la muestra, el 5.5% están infectadas por VIH

1.2. Incidencia Acumulada (o Riesgo)

La incidencia acumulada (o riesgo) de una enfermedad es la probabilidad de que la enfermedad ocurra durante un período de tiempo determinado. Se calcula utilizando la siguiente fórmula:

$$\text{Riesgo} = \text{Incidencia Acumulada} = \frac{\text{Número casos nuevos de enfermedad durante el periodo}}{\text{Número individuos libres de enfermedad al inicio del periodo}}$$

Supongamos que estamos interesados en determinar el riesgo de muerte al año del diagnóstico de una angina de pecho. Si seguimos durante un año a 2418 hombres con angina de pecho, de los que 456 se mueren al año del diagnóstico, el riesgo de muerte al año del diagnóstico de la enfermedad se calcularía como:

$$Riesgo = \frac{456}{2418} = 18.9\%$$

Entre los 2418 hombres de la muestra, el 18.9% se mueren al año del diagnóstico de la enfermedad.

*La **incidencia acumulada (o riesgo)** de una enfermedad es la probabilidad de que la enfermedad ocurra durante un período de tiempo determinado.*

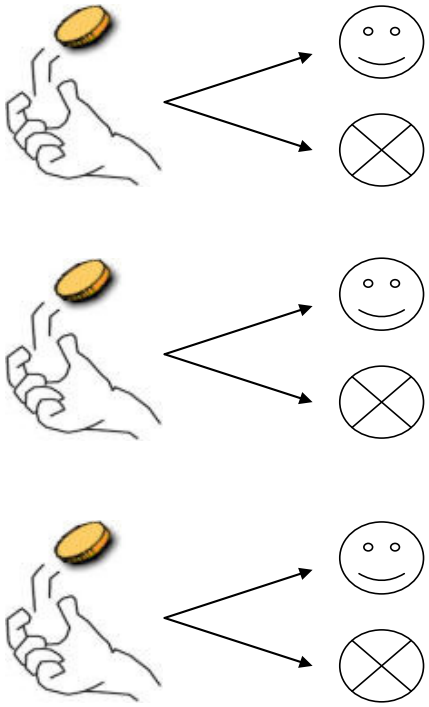
2. Distribución Binomial

Acabamos de ver cómo calcular la proporción de individuos que presentan la característica de interés a partir de los datos de la muestra. Sin embargo, el principal objetivo del Análisis Estadístico es obtener conclusiones sobre la proporción de individuos que presentan la característica de interés en la población, a partir de la información proporcionada por la muestra. Para poder hacer esto, en primer lugar necesitamos conocer la distribución en el muestreo de una proporción.

La distribución en el muestreo de una proporción se conoce como distribución Binomial. Para presentar la distribución Binomial, vamos a considerar el experimento de lanzar una moneda al aire. Al lanzar la moneda, podemos obtener únicamente 2 resultados: Cara o Cruz, cada uno de ellos con una probabilidad de 0.5. Dado que el resultado de lanzar una moneda al aire sólo puede ser Cara o Cruz, el resultado del lanzamiento es una variable dicotómica. Imaginemos que de los dos posibles resultados que podemos observar al lanzar la moneda, la característica de interés, la que estamos interesados en estudiar, es observar Cara.

Ahora repetimos el experimento de lanzar una moneda al aire 3 veces y observamos el Número de caras obtenidas en los 3 lanzamientos.

INDEPENDIENTES

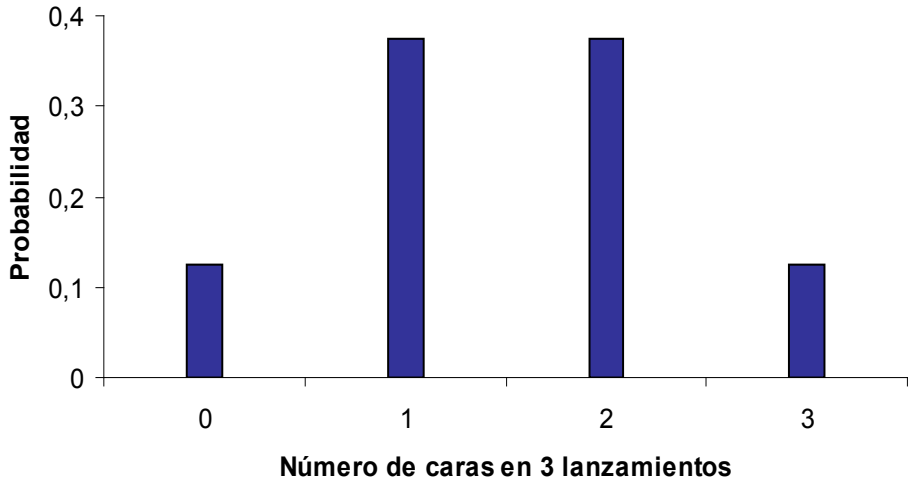


Si lanzamos una moneda al aire 3 veces, podemos observar 0 caras, 1, 2 ó 3 caras. De forma intuitiva, podemos calcular la Probabilidad de obtener 0, 1, 2 ó 3 caras al lanzar una moneda al aire 3 veces. Por ejemplo, obtenemos 0 caras si el resultado de cada uno de los 3 lanzamientos es una cruz. La probabilidad de obtener Cruz en el primer lanzamiento es 0.5, la misma que la de obtener Cruz en el segundo y en el tercer lanzamiento. Dado que el resultado obtenido en un lanzamiento es independiente del resultado obtenido en los otros lanzamientos, podemos calcular la Probabilidad de obtener 0 caras, o equivalentemente 3 cruces, como el producto de 0.5 por 0.5 por 0.5.

Para calcular la probabilidad de obtener 1 cara, en primer lugar nos planteamos qué situaciones pueden dar lugar a observar 1 cara en 3 lanzamientos. Hay 3 situaciones que darían lugar a observar 1 cara: (Cara, Cruz, Cruz), (Cruz, Cara, Cruz) y (Cruz, Cruz, Cara). Siguiendo el razonamiento anterior, la probabilidad de observar (Cara, Cruz, Cruz) sería $0.5x0.5x0.5 = (0.5)^3$. Del mismo modo, la probabilidad de observar (Cruz, Cara, Cruz) sería $(0.5)^3$ y la probabilidad de observar (Cruz, Cruz, Cara) sería $(0.5)^3$. Por lo tanto, como hay 3 posibles formas de observar 1 cara, cada una de ellas con probabilidad $(0.5)^3$, la probabilidad de observar 1 cara se calcularía como $3x(0.5)^3$. Siguiendo el mismo razonamiento, podemos calcular la probabilidad de observar 2 y 3 caras, respectivamente.

Nº caras	Nº situaciones	Probabilidad
0	XXX	$1 \cdot (0.5)^3 = 0.125$
1	CXX XCX XXC	$3 \cdot (0.5)^3 = 0.375$
2	CCX CXC XCC	$3 \cdot (0.5)^3 = 0.375$
3	CCC	$1 \cdot (0.5)^3 = 0.125$

Las probabilidades obtenidas pueden representarse gráficamente en lo que se conoce como Distribución de probabilidad Binomial (n = 3, p = 0.5), donde n es el número de lanzamientos y p la probabilidad de observar cara.



2.1. Fórmula general de la Distribución Binomial

La fórmula general para calcular la Probabilidad de observar exactamente *d* eventos en una muestra de n individuos, siendo p la probabilidad de que cada individuo experimente el evento es:

$$\Pr(d\ eventos) = \frac{n!}{d!(n-d)!}\pi^{d}(1-\pi)^{n-d}$$

La primera parte de la fórmula representa el número posible de situaciones en las que pueden ocurrir d eventos entre los n individuos observados, y la segunda parte de la fórmula representa la probabilidad de que cada una de las situaciones ocurra.

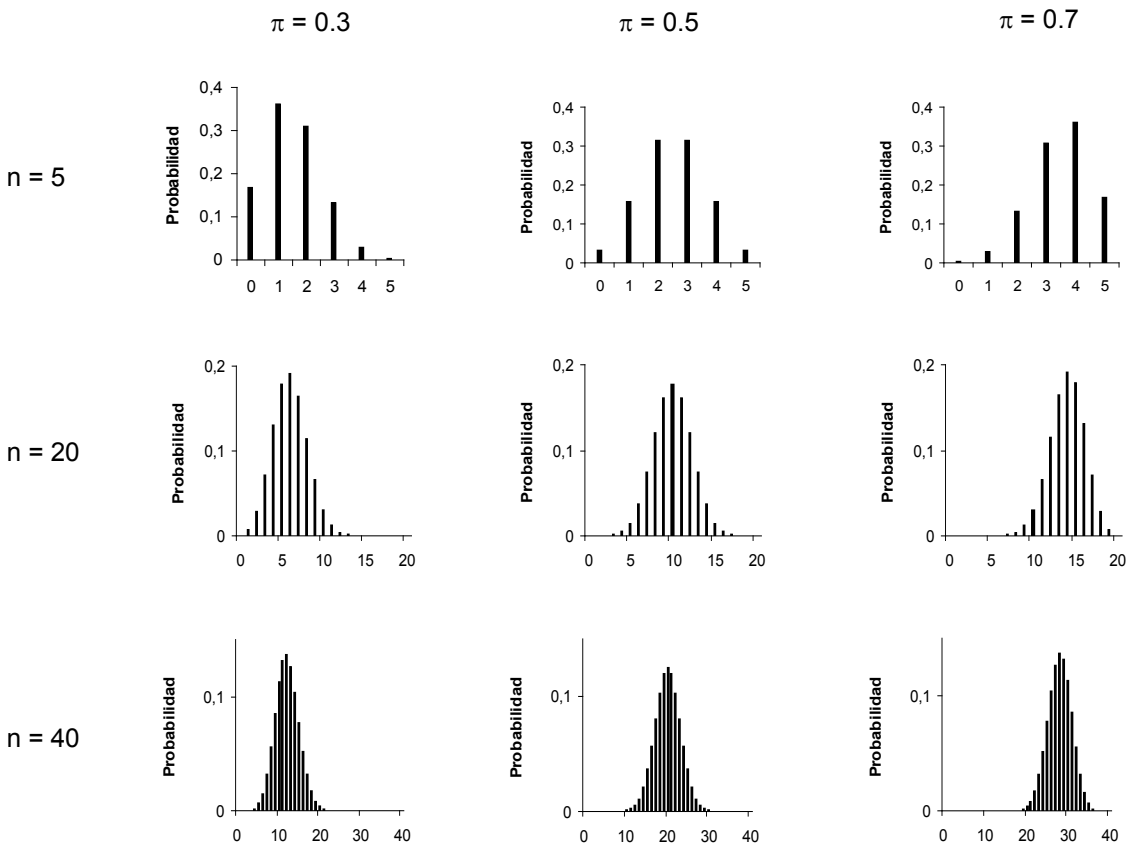
La exclamación (!) denota el factorial de un número. Por ejemplo, n! = n·(n-1) ·(n-2) ·(n-3) ·…·1. El factorial de 0 es 1.

En el experimento de lanzar una moneda al aire 3 veces, el número de caras obtenidas en los 3 lanzamientos sigue una distribución Binomial (n = 3, p = 0.5). Podemos calcular, por ejemplo, la Probabilidad de que ocurran 2 caras al lanzar una moneda al aire 3 veces de la siguiente forma:

$$\Pr(2\ caras) = \frac{n!}{d!(n-d)!}\pi^{d}(1-\pi)^{n-d} = \frac{3!}{2!(3-2)!}0.5^{2}(1-0.5)^{3-2} = 0.375$$

2.2.Forma de la Distribución Binomial

En la siguiente Figura se presentan ejemplos de distribuciones de probabilidad Binomiales para diferentes valores de n y p. Estas distribuciones se han ilustrado para *d*, el número de individuos que presentan la característica de interés en la muestra, pero pueden aplicarse a *p*, la proporción de individuos de la muestra que presentan la característica de interés. Por ejemplo, cuando el tamaño muestral (n) es 5, los posibles valores de *d* son 0, 1, 2, 3, 4 y 5. Equivalentemente, podríamos haber representado las proporciones *p* correspondientes (0, 0.2, 0.4, 0.6, 0.8 y 1) a los valores de *d* (0, 1, 2, 3, 4 y 5).



La distribución Binomial es la distribución en el muestreo para el número (o proporción) de individuos que presentan una determinada característica de interés. Por lo tanto, la media de la distribución en el muestreo es la media poblacional, y la desviación estándar de la distribución en el muestreo representa el error estándar. En la siguiente tabla se presenta la media y el error estándar de la distribución en el muestreo del número y proporción de individuos que presentan una determinada característica de interés:

*La **distribución Binomial** es la distribución en el muestreo para el número (o proporción) de individuos que presentan una determinada característica de interés.*

	Valor observado	Media poblacional	Error estándar
Número de individuos	*d*	$n\pi$	$\sqrt{n\pi(1-\pi)}$
Proporción de individuos	*p* = *d*/n	π	$\sqrt{(\pi(1-\pi)\ /n}$

En lo que sigue, nos basaremos en la distribución en el muestreo de la proporción de individuos que presentan una determinada característica de interés para calcular Intervalos de Confianza y

realizar Contrastes de Hipótesis para la proporción de individuos que presentan la característica de interés en la población a estudio.

*La **distribución Binomial se aproxima** a la **distribución Normal** conforme aumenta el tamaño de la muestra. La aproximación de la distribución Binomial a la distribución Normal nos permitirá calcular **Intervalos de Confianza** y realizar **Contrastes de Hipótesis** sobre la **Proporción de individuos** que presentan una determinada característica de interés en la población a estudio.*

2.3. Aproximación de la distribución Binomial a la distribución Normal

Tal y como hemos visto previamente, la distribución Binomial se aproxima a la distribución Normal conforme aumenta el tamaño de la muestra, n.

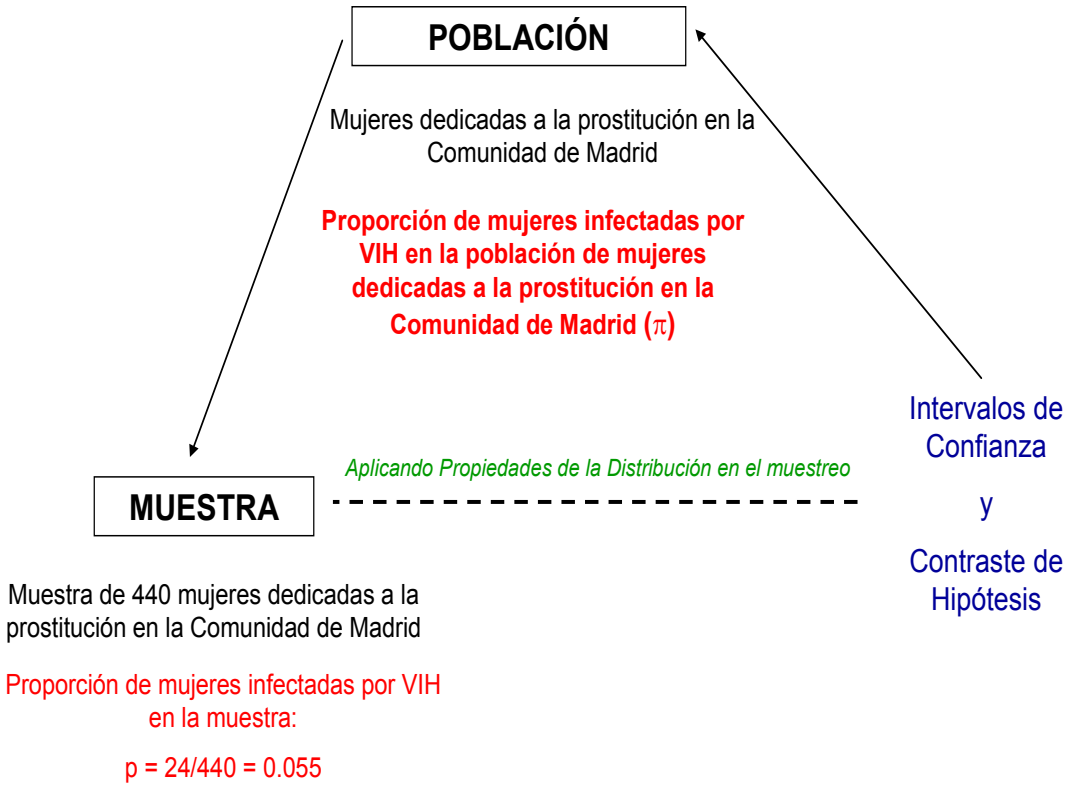
La aproximación de la Distribución Binomial a la Distribución Normal es una buena aproximación si tanto np como n- np toman un valor igual o superior a 10.

La aproximación de la Distribución Binomial a la Distribución Normal nos permitirá calcular Intervalos de Confianza y realizar Contrastes de Hipótesis sobre la Proporción de individuos que presentan una determinada característica de interés en la población a estudio.

3. Inferencia sobre una proporción

Supongamos que estamos interesados en determinar la proporción de mujeres infectadas por VIH entre las mujeres que se dedican a la prostitución en la Comunidad de Madrid. La población a estudio es el colectivo de mujeres que se dedican a la prostitución en la Comunidad de Madrid. El parámetro poblacional que estamos interesados en conocer es la proporción de mujeres infectadas por VIH. En lo que sigue, a la proporción de individuos que satisfacen una determinada característica de interés en la población lo denotaremos como π.

Por razones técnicas y financieras no podemos acceder al total de mujeres de la población a estudio. En su defecto, seleccionamos una muestra de 440 mujeres de la población, a las que les hacemos la prueba del VIH. Del total de las 440 mujeres, 24 están infectadas por VIH; por lo tanto, la proporción de mujeres infectadas por VIH en la muestra es p = 24/440 = 0.055.



Si hubiéramos seleccionado una muestra diferente, habríamos obtenido una proporción diferente de mujeres infectadas por VIH, debido a lo que se conoce como **Variación en el muestreo,** es decir, debido a la variabilidad que surge de observar muestras en lugar de poblaciones.

Para determinar la proporción de mujeres infectadas por VIH en el colectivo de mujeres que ejercen la prostitución en la Comunidad de Madrid (π), calculamos un Intervalo de Confianza para una Proporción.

Antes de esto, necesitamos conocer la distribución en el muestreo de una proporción. En los apartados 2.2 y 2.3 se ha mostrado que si el tamaño muestral, n, es lo suficientemente grande:

- La distribución en el muestreo de las proporciones es aproximadamente Normal
- La Media de la distribución en el muestreo de las proporciones es la Proporción poblacional (π)
- La desviación estándar de la distribución en el muestreo de

las proporciones, conocida como Error estándar, es:

$$EE(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Dado que la proporción poblacional, p, es desconocida, utilizaremos la proporción muestral para estimar el error estándar de las proporciones muestrales. Por lo tanto,

$$EE(p) = \sqrt{\frac{p(1-p)}{n}}$$

3.1. Intervalo de Confianza al 95% para una Proporción

La fórmula general para calcular un Intervalo de Confianza al 95% para un parámetro poblacional es:

$$IC_{95\%}(parámetro) = estimador \pm 1.96 \times EE(estimador)$$

Por lo tanto, un Intervalo de Confianza al 95% para la Proporción de individuos que presentan una determinada característica de interés en la población a estudio se obtendría como:

$$IC_{95\%}(\pi) = p \pm 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$

En nuestro ejemplo, un Intervalo de Confianza al 95% para la Proporción de mujeres infectadas por VIH en el colectivo de mujeres que ejercen la prostitución en la Comunidad de Madrid se calcularía como:

Proporción de mujeres infectadas por VIH en la muestra: p = 24/440 = 0.055

Error estándar estimado:

$$EE(p) = \sqrt{\frac{0.055(1-0.055)}{440}} = \sqrt{\frac{0.052}{440}} = \sqrt{0.00012} = 0.011$$

$$IC_{95\%}(\pi) = p \pm 1.96 \cdot \sqrt{\frac{p(1-p)}{n}} = 0.055 \pm (1.96 \cdot 0.011) = (0.033\ ;\ 0.077)$$

Estamos seguros al 95% de que la Proporción de mujeres infectadas por VIH en el colectivo de mujeres que se dedican a la prostitución en la Comunidad de Madrid está entre 0.033 y 0.077 (o, equivalentemente, entre el 3.3% y el 7.7%).

Muestras pequeñas

Los métodos descritos para calcular un Intervalo de Confianza para la Proporción de individuos que presentan una determinada característica de interés en la población se basan en muestras grandes.

No hay una regla fácil y rápida para cuando el tamaño de la muestra no es suficientemente grande. Si la proporción de individuos que presentan una determinada característica de interés en la muestra es menor de 0.05 y el numerador (número de individuos con la característica de interés), es 5 ó más, los métodos descritos previamente pueden utilizarse sin problema. En caso contrario, necesitamos utilizar "Métodos Exactos" basados en la Distribución Binomial para calcular Intervalos de Confianza y realizar Contrastes de Hipótesis.

*Si el **tamaño de la muestra no es suficientemente grande**, es necesario utilizar **"Métodos Exactos"** basados en la Distribución Binomial para calcular Intervalos de Confianza y realizar Contrastes de Hipótesis sobre la Proporción de individuos que presentan una determinada característica de interés en la población.*

4. Comparación de dos proporciones

En la Investigación Médica, son numerosas las ocasiones en las que el interés se centra en determinar si la proporción de individuos que presentan una determinada característica de interés es igual en dos grupos diferentes de individuos, expuestos y no expuestos a una determinada variable de exposición. En este punto, se presentan los métodos a utilizar para estudiar la asociación entre una variable de exposición dicotómica y una variable de interés dicotómica.

4.1 Tabla de contingencia 2x2

Supongamos que estamos interesados en estudiar si existe una asociación estadísticamente significativa entre el uso de drogas por vía parenteral y la infección por VIH, es decir, si la proporción de mujeres infectadas por VIH es la misma entre las mujeres que no usan drogas por vía parenteral que entre las mujeres que si que usan drogas por vía parenteral.

*De forma descriptiva, la relación entre dos variables dicotómicas puede examinarse mediante una **Tabla de contingencia 2x2**. En las filas se representan las categorías de la variable de exposición, en las columnas las categorías de la variable de interés, y en cada intersección se representa el número de veces que dicho par de valores se ha presentado conjuntamente.*

De forma descriptiva, la relación entre dos variables dicotómicas puede examinarse mediante una Tabla de contingencia 2x2, una tabla que reúne conjuntamente la información de las dos variables, tal y como se muestra a continuación:

		Variable de interés		
		No	Si	Total
Variable de Exposición	No	h_0	d_0	n_0
	Si	h_1	d_1	n_1
Total		h	d	n

Por convención, en las filas se representan las categorías de la variable de exposición y en las columnas se representan las categorías de la variable de interés. En cada intersección de un valor de la variable por filas y otro valor de la variable por columnas se representa el número de veces que dicho par de valores se ha presentado conjuntamente.

En nuestro ejemplo, supongamos que de las 440 mujeres incluidas en la muestra, 391 no usan drogas y 49 usan drogas por vía parenteral. Un total de 24 mujeres están infectadas por VIH, 6 en las mujeres que no usan drogas y 18 en las que usan drogas por vía parenteral. A partir de estos datos, podemos construir una Tabla de contingencia, tal y como se muestra a continuación:

		Infección por VIH		
		No	Si	Total
Uso drogas por vía parenteral	No	385	6	391
	Si	31	18	49
Total		416	24	440

La interpretabilidad de las Tablas de contingencia puede mejorarse incluyendo porcentajes. Los porcentajes pueden calcularse respecto a la variable representada en las filas, respecto a la variable representada en las columnas, o respecto al total. Como recomendación, los porcentajes deberían calcularse respecto a la variable de exposición. Por lo tanto, si la variable de exposición se representa en las filas, los porcentajes deberían calcularse por filas, tal y como se muestra a continuación:

		Infección por VIH		
		No	Si	Total
Uso drogas por vía parenteral	No	385 (98.5%)	6 (1.5%)	391 (100.0%)
	Si	31 (63.3%)	18 (36.7%)	49 (100.0%)
Total		416 (94.5%)	24 (5.5%)	440 (100.0%)

Los resultados del análisis descriptivo muestran que el porcentaje de mujeres infectadas por el VIH es del 1.5% entre las mujeres que no usan drogas por vía parenteral y del 36.7% entre las mujeres que si que usan drogas por vía parenteral.

4.2.Test chi-cuadrado de asociación para Tablas 2x2

A partir de los datos de la muestra, observamos que la proporción de mujeres infectadas por VIH es considerablemente mayor entre las mujeres que usan drogas por vía parenteral que entre las mujeres que no usan drogas. Pero, ¿la mayor proporción de mujeres infectadas por VIH observada entre las mujeres que usan drogas por vía parenteral puede explicarse por azar? Para dar respuesta a esta pregunta, necesitamos realizar un Contraste de Hipótesis, denominado Test chi-cuadrado de asociación.

En primer lugar, definimos la Hipótesis Nula y la Hipótesis Alternativa de la siguiente forma:

Hipótesis Nula (H_0):

En la población de mujeres que se dedican a la prostitución en la Comunidad de Madrid, no existe una asociación estadísticamente significativa entre el Uso de drogas por vía parenteral y la Infección por VIH; es decir, la proporción de mujeres infectadas por VIH es la misma entre las mujeres que no usan drogas por vía parenteral que entre las mujeres que si que usan drogas por vía parenteral

Hipótesis Alternativa (H_1):

En la población de mujeres que se dedican a la prostitución en la Comunidad de Madrid, existe una asociación estadísticamente significativa entre el Uso de drogas por vía parenteral y la Infección por VIH; es decir, la proporción de mujeres infectadas por VIH es diferente entre las mujeres que no usan drogas por

vía parenteral que entre las mujeres que si que usan drogas por vía parenteral

A continuación, calculamos el test estadístico. El test estadístico chi-cuadrado se basa en la comparación de las frecuencias observadas en cada una de las celdas de la Tabla de Contingencia con las frecuencias que se esperarían si la Hipótesis Nula fuera cierta. El cálculo de las frecuencias esperadas se lleva a cabo mediante la siguiente fórmula:

$$Frecuencia\ esperada = \frac{Total\ Fila \times Total\ Columna}{Total}$$

Por ejemplo, la Frecuencia esperada de la celda (No usa drogas; No infección por VIH) se calcularía como

$$\frac{391 \times 416}{440} = 369.7$$

En la siguiente Tabla, se muestran las frecuencias observadas y las frecuencias esperadas en cada una de las celdas de la Tabla de contingencia:

		Infección por VIH		
		No	Si	Total
Uso drogas por vía parenteral	No	385 369.7	6 21.3	391
	Si	31 46.3	18 2.7	49
Total		416	24	440

A continuación, calculamos el test estadístico chi cuadrado mediante la siguiente fórmula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O: Frecuencia observada en cada celda

E: Frecuencia esperada en cada celda

Σ: Sumatorio a través de todas las celdas

El valor del estadístico χ^2 será pequeño si las diferencias entre los valores observados y esperados son insignificantes. En este caso, los datos observados son parecidos a los que esperaríamos si la Hipótesis Nula fuera cierta, indicando que no hay evidencia en contra de la Hipótesis Nula de no asociación. El valor del estadístico chi-cuadrado será grande si hay diferencias

considerables entre los valores observados y esperados. En este caso, los datos observados no son los que esperaríamos si la Hipótesis Nula fuera cierta, indicando que los datos proporcionan evidencia en contra de la Hipótesis Nula.

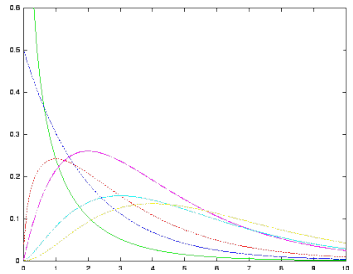
Bajo la Hipótesis Nula, el estadístico chi-cuadrado sigue una distribución chi-cuadrado con 1 grado de libertad.

Distribución chi-cuadrado

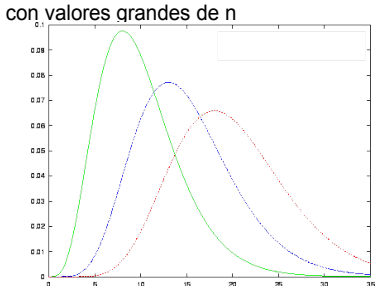
- Distribución continua, asociada principalmente a los tests chi-cuadrado
- Determinada por un parámetro conocido como Grados de libertad

  χ_n^2 es una distribución chi-cuadrado con n grados de libertad
- El rango de la distribución chi-cuadrado es el eje real positivo: (0, ∞)

Función de densidad de una distribución χ_n^2 con valores pequeños de n



Función de densidad de una distribución χ_n^2 con valores grandes de n



En nuestro ejemplo, el valor del estadístico chi-cuadrado se calcularía como:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(385-369.7)^2}{369.7} + \frac{(6-21.3)^2}{21.3} + \frac{(31-46.3)^2}{46.3} + \frac{(18-2.7)^2}{2.7} = 104.62$$

El p-valor se calcularía como la probabilidad de observar una diferencia entre las frecuencias esperadas y observadas como la obtenida o más extrema, si la Hipótesis Nula fuera cierta; es decir,

$$p-valor = \Pr(\chi_1^2 \geq 104.62) < 0.001$$

El p-valor del test chi-cuadrado de asociación es <0.001. Los datos presentan evidencia estadística suficiente para rechazar la Hipótesis Nula. Existe una asociación estadísticamente significativa entre el uso de drogas por vía parenteral y la infección por VIH, es decir, la proporción de mujeres infectadas por VIH es diferente entre las mujeres que no usan drogas por vía parenteral que entre las mujeres que si que usan drogas por vía parenteral.

El Test chi-cuadrado de asociación para Tablas 2x2 permite determinar si existe una asociación estadísticamente significativa entre dos variables dicotómicas.

Hay tres medidas que nos permiten cuantificar la magnitud de la asociación entre una variable de exposición dicotómica y una variable de interés dicotómica:
- Diferencia de proporciones (DP)
- Razón de proporciones (RP)
- Odds ratio (OR)

4.3. Magnitud de la asociación para Tablas 2x2

Una vez que hemos detectado la existencia de una asociación entre el uso de drogas por vía parenteral y la infección por VIH, el siguiente paso consiste en cuantificar la magnitud de la asociación. Hay tres medidas que nos permiten cuantificar la magnitud de la asociación:

- Diferencia de proporciones (DP)
- Razón de proporciones (RP)
- Odds ratio (OR)

➢ Diferencia de proporciones

La diferencia de proporciones se define como:

$$Diferencia\ \Pr oporciones = p_1 - p_0$$

donde

p_0 (proporción de individuos que presentan la característica de interés en el grupo de No Expuestos) se calcula como:

$$p_0 = \frac{d_0}{n_0}$$

p_1 (proporción de individuos que presentan la característica de interés en el grupo de Expuestos) se calcula como:

$$p_1 = \frac{d_1}{n_1}$$

En nuestro ejemplo, si consideramos como No expuestas a las mujeres que no consumen drogas por vía parenteral, y como Expuestas a las mujeres que si que consumen drogas por vía parenteral, la diferencia entre la proporción de mujeres infectadas por VIH entre las que usan drogas por vía parenteral y las que no usan drogas por vía parenteral se calcularía como:

		Infección por VIH		
		No	Si	Total
Uso drogas por vía parenteral	No	385	6	391
	Si	31	18	49
Total		416	24	440

Proporción de mujeres infectadas por VIH entre las que no usan drogas por vía parenteral:

$$p_0 = \frac{d_0}{n_0} = \frac{6}{391} = 0.015$$

Proporción de mujeres infectadas por VIH entre las que usan drogas por vía parenteral:

$$p_1 = \frac{d_1}{n_1} = \frac{18}{49} = 0.367$$

$$Diferencia\ \text{Pr}oporciones = p_1 - p_0 = 0.367 - 0.015 = 0.352$$

En las 440 mujeres de la muestra, la diferencia en la proporción de mujeres infectadas por VIH entre las mujeres que usan drogas por vía parenteral y las que no usan drogas por vía parenteral es 0.352.

Una vez hemos cuantificado la magnitud de la asociación en los individuos de la muestra, calculamos un Intervalo de Confianza para la Diferencia de Proporciones, algo que nos permite cuantificar la magnitud de la asociación en la población de la que se extrajo la muestra. Para ello, en primer lugar, necesitamos conocer la distribución en el muestreo de la Diferencia de dos Proporciones.

Una vez hemos cuantificado la magnitud de la asociación en los individuos de la muestra, calculamos un ***Intervalo de Confianza para la Diferencia de Proporciones****, algo que nos permite cuantificar la magnitud de la asociación en la población de la que se extrajo la muestra.*

Distribución en el muestreo de la Diferencia de dos Proporciones

Se puede demostrar que la distribución en el muestreo de la Diferencia entre dos Proporciones tiene las siguientes propiedades:

1. La distribución en el muestreo de la diferencia de dos proporciones muestrales es aproximadamente Normal
2. La media de la distribución en el muestreo de la diferencia de dos proporciones es la diferencia entre las dos proporciones poblacionales $(\pi_1 - \pi_0)$
3. El error estándar de la diferencia entre las dos proporciones muestrales es una combinación de los errores estándar de las proporciones individuales:

$$EE(p_1 - p_0) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

Intervalo de Confianza al 95% para la Diferencia entre dos Proporciones

La fórmula general para calcular un Intervalo de Confianza al 95% para un parámetro poblacional es:

$$IC_{95\%}(parámetro) = estimador \pm 1.96 \times EE(estimador)$$

Por lo tanto, un Intervalo de Confianza al 95% para la Diferencia entre dos Proporciones se obtendría como:

$$IC_{95\%}(\pi_1 - \pi_0) = (p_1 - p_0) \pm 1.96 \times EE(p_1 - p_0) = (p_1 - p_0) \pm 1.96 \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

En nuestro ejemplo, un Intervalo de Confianza al 95% para la diferencia en la proporción de mujeres infectadas por VIH entre las mujeres que usan drogas por vía parenteral y las que no usan drogas por vía parenteral viene dado por:

$$IC_{95\%}(\pi_1 - \pi_0) = (p_1 - p_0) \pm 1.96 \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

$$= (0.367 - 0.015) \pm 1.96 \times \sqrt{\frac{0.367(1-0.367)}{49} + \frac{0.015(1-0.015)}{391}} =$$

$$= 0.352 \pm 1.96 \times 0.069 = (0.22;\ 0.49)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la diferencia en la proporción de mujeres infectadas por VIH entre las mujeres que usan drogas por vía parenteral y las que no usan drogas por vía parenteral está entre 0.22 y 0.49.

Como el Intervalo de Confianza para la diferencia de proporciones no incluye al 0, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la proporción de mujeres infectadas por VIH entre las mujeres que no usan drogas NO es la misma que la proporción de mujeres infectadas por VIH entre las mujeres que usan drogas por vía parenteral.

La conclusión a la que llegamos con el Intervalo de Confianza para la Diferencia de Proporciones es la misma a la que habíamos llegado previamente mediante el test chi cuadrado de asociación; es decir, existe una asociación estadísticamente significativa

entre el uso de drogas por vía parenteral y la infección por VIH. La ventaja del Intervalo de Confianza frente al Contraste de Hipótesis es que el Intervalo de Confianza no sólo nos permite saber si hay asociación entre la variable de exposición y la variable de interés, sino que también nos permite cuantificar la magnitud de la asociación en la población.

La conclusión a la que llegamos con el Intervalo de Confianza para la Diferencia de Proporciones es la misma a la que llegamos mediante el test chi cuadrado de asociación. La ventaja del Intervalo de Confianza frente al Contraste de Hipótesis es que el Intervalo de Confianza no sólo nos permite saber si hay asociación entre la variable de exposición y la variable de interés, sino que también nos permite cuantificar la magnitud de la asociación en la población.

- **Razón de proporciones**

La razón de proporciones se define como:

$$RP = \frac{p_1}{p_0}$$

donde

p_0 (proporción de individuos que presentan la característica de interés en el grupo de No Expuestos) se calcula como:

$$p_0 = \frac{d_0}{n_0}$$

p_1 (proporción de individuos que presentan la característica de interés en el grupo de Expuestos) se calcula como:

$$p_1 = \frac{d_1}{n_1}$$

En nuestro ejemplo, la razón entre la proporción de mujeres infectadas por VIH en las mujeres que usan drogas y la proporción de mujeres infectadas por VIH en las mujeres que no usan drogas se calcularía como:

		Infección por VIH		
		No	Si	Total
Uso drogas por vía parenteral	No	385	6	391
	Si	31	18	49
Total		416	24	440

Proporción de mujeres infectadas por VIH entre las que no usan drogas por vía parenteral:

$$p_0 = \frac{d_0}{n_0} = \frac{6}{391} = 0.015$$

Proporción de mujeres infectadas por VIH entre las que usan drogas por vía parenteral:

$$p_1 = \frac{d_1}{n_1} = \frac{18}{49} = 0.367$$

$$RP = \frac{p_1}{p_0} = \frac{0.367}{0.015} = 23.73$$

Interpretación

La razón de proporciones siempre es un número positivo:

RP = 1	La proporción de individuos que presentan la característica de interés es la misma en el grupo de individuos expuestos que en el grupo de individuos no expuestos o, equivalentemente, no hay asociación entre la variable de exposición y la variable de interés
RP > 1	La proporción de individuos que presentan la característica de interés es mayor entre los expuestos que entre los no expuestos, sugiriendo que la exposición es un factor de riesgo para presentar la característica de interés
RP < 1	La proporción de individuos que presentan la característica de interés es menor entre los expuestos que entre los no expuestos, sugiriendo que la exposición es protectora para presentar la característica de interés

Conforme la razón de proporciones se aleja de 1, mayor es la asociación entre la variable de exposición y la variable de interés.

En nuestro ejemplo, el valor de RP en las 440 mujeres de la muestra es 23.73; la proporción de mujeres infectadas por VIH entre las mujeres que usan drogas es 23.73 veces superior que la proporción de mujeres infectadas por VIH entre las mujeres que no usan drogas por vía parenteral. Es decir, el consumo de drogas por vía parenteral multiplica por 23.73 la proporción de mujeres infectadas por VIH.

En las situaciones en las que la proporción calculada se refiere a un riesgo (o incidencia acumulada), la razón de proporciones se conoce con el nombre de Riesgo Relativo (RR).

Intervalo de Confianza al 95% para la Razón de Proporciones

Una vez hemos calculado la Razón de Proporciones en los individuos de la muestra, el siguiente paso consiste en calcular

La razón de proporciones se define como:

$$RP = \frac{p_1}{p_0}$$

Donde p_0 es la proporción de individuos que presentan la característica de interés en el grupo de No Expuestos, y p_1 es la proporción de individuos que presentan la característica de interés en el grupo de Expuestos

un Intervalo de Confianza al 95% para la Razón de Proporciones en la población de la que se extrajeron las muestras.

El procedimiento general utilizado hasta ahora para el cálculo de Intervalos de Confianza al 95% ha sido:

$$IC_{95\%}(parámetro) = estimador \pm 1.96 \times EE(estimador)$$

En el caso de medidas relativas o de razón (ej. Razón de Proporciones o Odds Ratio), este procedimiento puede ser problemático, pudiéndose observar límites inferiores de los Intervalos de Confianza negativos, a pesar de que tanto la Razón de Proporciones como la Odds Ratio siempre tomarán valores positivos. Esto puede ocurrir si el error estándar es grande, y/o si la medida relativa o de razón utilizada toma un valor próximo a 0.

Para superar este problema, utilizaremos el siguiente procedimiento para calcular un Intervalo de Confianza al 95% para la Razón de Proporciones:

1. Calculamos el logaritmo de la Razón de Proporciones, y su error estándar (Método Delta):

$$EE(\log RP) = \sqrt{\frac{1}{d_1} + \frac{1}{n_1} + \frac{1}{d_0} + \frac{1}{n_0}}$$

2. Calculamos un Intervalo de Confianza al 95% para el logRP de la forma habitual:

$$IC_{95\%}(\log RP) = \log RP \pm (1.96 \times EE(\log RP))$$

3. Tomamos la exponencial de los límites del Intervalo de Confianza obtenido para obtener un Intervalo de Confianza de la Razón de Proporciones

Siguiendo el procedimiento anterior, obtenemos la fórmula para calcular un Intervalo de Confianza al 95% para la Razón de Proporciones:

$$IC_{95\%}(RP) = \left[\frac{RP}{\exp(1.96 \times EE(\log RP))} ; RP \times \exp(1.96 \times EE(\log RP)) \right]$$

En nuestro ejemplo,

		Infección por VIH		
		No	Si	Total
Uso drogas por vía parenteral	No	385	6	391
	Si	31	18	49
Total		416	24	440

$$RP = \frac{p_1}{p_0} = \frac{0.367}{0.015} = 23.73$$

$$\log RP = \log(23.73) = 3.17$$

$$EE(\log RP) = \sqrt{\frac{1}{d_1} + \frac{1}{n_1} + \frac{1}{d_0} + \frac{1}{n_0}} = \sqrt{\frac{1}{18} + \frac{1}{49} + \frac{1}{6} + \frac{1}{391}} = 0.50$$

$$IC_{95\%}(RP) = \left[\frac{23.73}{\exp(1.96 \times 050)}; 23.73 \times \exp(1.96 \times 0.50)) \right] = (8.91; 63.23)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la proporción de mujeres infectadas por VIH entre las mujeres que usan drogas es entre 8.91 y 63.23 veces superior que la proporción de mujeres infectadas por VIH entre las mujeres que no usan drogas por vía parenteral. Es decir, el consumo de drogas por vía parenteral multiplica entre 8.91 y 63.23 veces la proporción de mujeres infectadas por VIH.

Como el Intervalo de Confianza para la razón de proporciones no incluye al 1, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la proporción de mujeres infectadas por VIH entre las mujeres que no usan drogas NO es la misma que la proporción de mujeres infectadas por VIH entre las mujeres que usan drogas por vía parenteral.

Consideraciones sobre la Razón de Proporciones

La Razón de Proporciones es una medida de asociación fácil de interpretar. Los análisis basados en Razones de Proporciones descritos hasta el momento son directos y sencillos de llevar a cabo.

Sin embargo, análisis más complicados para el estudio de asociaciones entre variables de exposición y variables de interés dicotómicas, NO se basan en el cálculo de Razones de Proporciones. Los análisis complejos se basan en una medida de asociación conocida como Odds Ratio (OR).

*La **Odds Ratio** se define como razón entre la odds de presentar la característica de interés en el grupo de individuos expuestos y la odds de presentar la característica de interés en el grupo de individuos no expuestos.*

- **Odds ratio**

La Odds Ratio se define como la razón entre la odds de presentar la característica de interés en el grupo de individuos expuestos y la odds de presentar la característica de interés en el grupo de individuos no expuestos.

La Odds se define como la proporción de individuos que presentan la característica de interés entre la proporción de individuos que NO presentan la característica de interés:

$$Odds = \frac{p}{1-p} = \frac{d/n}{(1-d/n)} = \frac{d/n}{h/n} = d/h$$

$$Odds\ Ratio = OR = \frac{odds\ en\ \text{exp}uestos}{odds\ en\ no\ \text{exp}uestos} = \frac{d_1/h_1}{d_0/h_0} = \frac{d_1 \times h_0}{d_0 \times h_1}$$

Donde d_0: número de individuos que presentan la característica de interés en el grupo de no expuestos

d_1: número de individuos que presentan la característica de interés en el grupo de Expuestos

h_0: Número de individuos que NO presentan la característica de interés en el grupo de no expuestos

h_1: número de individuos que NO presentan la característica de interés en el grupo de expuestos

En nuestro ejemplo,

		Infección por VIH		
		No	Si	Total
Uso drogas por vía parenteral	No	385	6	391
	Si	31	18	49
Total		416	24	440

La odds de infección por VIH entre las mujeres que usan drogas se calcularía como:

$$odds_{usan\ drogas} = d_1 / h_1 = 18/31 = 0.581$$

La odds de infección por VIH entre las mujeres que no usan drogas por vía parenteral se calcularía como:

$$odds_{no\ usan\ drogas} = d_0 / h_0 = 6/385 = 0.016$$

Por lo tanto, la OR de la asociación entre el uso de drogas por vía parenteral y la infección por VIH se calcularía como

$$OR_{uso\ drogas\ vs.\ no\ uso\ drogas} = \frac{odds_{usan drogas}}{odds_{no\ usan drogas}} = \frac{0.581}{0.016} = 36.31$$

Interpretación

La Odds Ratio (OR) siempre es un número positivo:

OR = 1	La odds de presentar la característica de interés, y por lo tanto, la proporción de individuos que presentan la característica de interés es la misma en el grupo de individuos expuestos que en el grupo de individuos no expuestos; equivalentemente, no hay asociación entre la variable de exposición y la variable de interés
OR > 1	La odds de presentar la característica de interés es mayor entre el grupo de individuos expuestos que entre el grupo de individuos no expuestos, sugiriendo que la exposición es un factor de riesgo para presentar la característica de interés
OR < 1	La odds de presentar la característica de interés es menor entre el grupo de individuos expuestos que entre el grupo de individuos no expuestos, sugiriendo que la exposición al factor es protectora para presentar la característica de interés

En nuestro ejemplo, el valor de la OR en las 440 mujeres de la muestra es 36.31; la odds de infección por VIH entre las mujeres que usan drogas es 36.31 veces superior que la odds de infección por VIH entre las mujeres que no usan drogas por vía parenteral. Es decir, el consumo de drogas por vía parenteral multiplica por 36.31 la odds de infección por VIH en las mujeres dedicadas a la prostitución.

Intervalo de Confianza al 95% para la Odds Ratio

Una vez hemos calculado la Odds Ratio de la asociación entre el uso de drogas por vía parenteral y la infección por VIH en los individuos de la muestra, el siguiente paso consiste en calcular un Intervalo de Confianza al 95% para la Odds Ratio en la población de mujeres dedicadas a la prostitución en la comunidad de Madrid.

El procedimiento para calcular un Intervalo de Confianza al 95% para la Odds Ratio (OR) es el siguiente:

1. Calculamos el logaritmo de la Odds Ratio, y su error estándar (fórmula de Woolf):

$$EE(\log OR) = \sqrt{\frac{1}{d_1} + \frac{1}{h_1} + \frac{1}{d_0} + \frac{1}{h_0}}$$

2. Calculamos un Intervalo de Confianza al 95% para el logOR de la forma habitual:

$$IC_{95\%}(\log OR) = \log OR \pm (1.96 \times EE(\log OR))$$

3. Tomamos la exponencial de los límites del Intervalo de Confianza obtenido para obtener un Intervalo de Confianza de la Odds Ratio (OR)

Siguiendo el procedimiento anterior, obtenemos la fórmula para calcular un Intervalo de Confianza al 95% para la Odds Ratio:

$$IC_{95\%}(OR) = \left[\frac{OR}{\exp(1.96 \times EE(\log OR))}; OR \times \exp(1.96 \times EE(\log OR)) \right]$$

En nuestro ejemplo,

$$OR = \frac{0.581}{0.016} = 36.31$$

$$\log OR = \log(36.31) = 3.59$$

$$EE(\log OR) = \sqrt{\frac{1}{d_1} + \frac{1}{h_1} + \frac{1}{d_0} + \frac{1}{h_0}} = \sqrt{\frac{1}{18} + \frac{1}{31} + \frac{1}{6} + \frac{1}{385}} = 0.51$$

$$IC_{95\%}(OR) = \left[\frac{OR}{\exp(1.96 \times EE(\log OR))}; OR \times \exp(1.96 \times EE(\log OR)) \right] =$$

$$= \left[\frac{36.31}{\exp(1.96 \times 0.51))}; 36.31 \times \exp(1.96 \times 0.51)) \right] = (13.36; 98.66)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la odds de infección por VIH entre las mujeres que usan drogas es entre 13.36 y 98.66 veces superior que la odds de infección por VIH entre las mujeres que no usan drogas por vía parenteral. Es decir, el consumo de drogas por vía parenteral multiplica entre 13.36 y 98.66 veces la odds de infección por VIH.

Como el Intervalo de Confianza para la odds ratio no incluye al 1, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la odds (y, por tanto, la proporción) de mujeres infectadas por VIH entre las mujeres que no usan drogas NO es la misma que la odds (y, por tanto, la proporción) de mujeres infectadas por VIH entre las mujeres que usan drogas por vía parenteral.

Consideraciones sobre la Odds Ratio

La Odds Ratio (OR) es siempre más lejana de 1 que la correspondiente Razón de Proporciones (RP). Si la RP es mayor que 1, entonces OR > RP. Si, por el contrario, RP<1, entonces OR<RP.

En el caso de eventos raros (la probabilidad de que la característica de interés NO ocurra es próxima a 1), la OR es aproximadamente igual a la RP, ya que la odds es aproximadamente igual que la proporción.

La Odds Ratio para la ocurrencia del evento es el inverso de la Odds Ratio para la no ocurrencia del evento.

La Odds Ratio para la exposición, es decir, la odds del evento en los expuestos dividida por la odds del evento en los no expuestos, es IGUAL a la Odds Ratio para el evento, esto es, la odds de la exposición en los individuos que experimentan el evento dividida por la odds de la exposición en los individuos que no experimentan el evento.

Razones para el uso de Odds Ratios

En la literatura médica reciente, los análisis estadísticos de variables de interés dicotómicas se basan casi siempre en Odds Ratios, aunque su interpretación es menos intuitiva que la de la

Diferencia de Proporciones o la Razón de Proporciones. Esto es así por 3 razones:

- Si el evento de interés es raro, la OR es igual a la RP. Esto es porque la odds de ocurrencia del evento raro es numéricamente equivalentemente a su riesgo. Análisis basados en OR dan los mismos resultados que los análisis basados en RP
- Cuando el evento de interés es común, los análisis basados en RP, particularmente aquéllos que examinan el efecto de más de una exposición, pueden causar problemas computacionales y son difíciles de interpretar. Estos problemas no ocurren con métodos basados en OR
- Para OR, las conclusiones son idénticas, independientemente de si consideramos nuestro evento de interés como la ocurrencia de un evento, o la ausencia del evento.

Además, las OR son las medidas de asociación utilizadas en los estudios de casos y controles.

5. Comparación de más de dos proporciones

En la Investigación Médica, son numerosas las ocasiones en las que el interés se centra en determinar si la proporción de individuos que presentan una determinada característica de interés es igual o diferente en más de dos grupos diferentes de individuos. En este punto, extendemos los métodos presentados previamente para el estudio de la asociación entre una variable de exposición con más de 2 categorías y una variable de interés dicotómica.

5.1. Tabla de contingencia rx2

Supongamos que estamos interesados en determinar si existe una asociación estadísticamente significativa entre la Edad, categorizada en 3 grupos (<25, 25-29, ≥30 años) y la Infección por VIH; es decir, si la proporción de mujeres infectadas por VIH varía en función de la Edad.

A partir de la información proporcionada por los datos de la muestra, sabemos que de las 440 mujeres del estudio, 154 son

menores de 25 años, 126 tienen entre 25 y 29 y 160 tienen 30 años ó más. De las 24 mujeres infectadas por VIH, 17 tienen menos de 25 años, 4 entre 25 y 29 y 3 tienen 3 años ó más.

A partir de esta información, construimos una Tabla de contingencia:

		Infección por VIH		
		No	Si	Total
Edad (categorizada)	<25	137	17	154
	25-29	122	4	126
	≥30	157	3	160
Total		416	24	440

Como se ha comentado en el apartado anterior, podemos mejorar la interpretabilidad de la Tabla de contingencia calculando los porcentajes por filas, es decir, calculando la proporción de mujeres infectadas por VIH en las 3 categorías de Edad:

		Infección por VIH		
		No	Si	Total
Edad (categorizada)	<25	137 (89.0%)	17 (11.0%)	154
	25-29	122 (96.8%)	4 (3.2%)	126
	≥30	157 (98.1%)	3 (1.0%)	160
Total		416 (94.5%)	24 (5.5%)	

El porcentaje de mujeres infectadas por VIH es: 11.0% en las mujeres menores de 25 años, 3.2% en las mujeres entre 25 y 29 años, y 1.9% en las mujeres de 30 años ó más. A la vista del análisis descriptivo, parece que existe una relación inversa entre la edad (categorizada) y el porcentaje de mujeres infectadas por VIH: a mayor edad, menor porcentaje de mujeres infectadas por VIH.

5.2. Test chi cuadrado de asociación para Tablas rx2

A partir de los datos de la muestra, hemos observado que la proporción de mujeres infectadas por VIH disminuye conforme aumenta la edad. Pero, la mayor proporción de mujeres infectadas por VIH observada entre las mujeres más jóvenes

puede ser explicada por azar? Para dar respuesta a esta pregunta, realizamos un Contraste de Hipótesis, denominado Test chi-cuadrado de asociación.

En primer lugar, definimos la Hipótesis Nula y la Hipótesis Alternativa de la siguiente forma:

Una Tabla de contingencia rx2 permite representar conjuntamente la información proporcionada por una variable de exposición con más de 2 categorías y una variable de interés dicotómica

Hipótesis Nula (H_0):

En la población de mujeres que se dedican a la prostitución en la Comunidad de Madrid, no existe una asociación estadísticamente significativa entre la Edad (categorizada) y la Infección por VIH; es decir, la proporción de mujeres infectadas por VIH es la misma entre las mujeres <25 años, las mujeres entre 25 y 29 años, y entre las mujeres de 30 años ó más.

Hipótesis Alternativa (H_1):

En la población de mujeres que se dedican a la prostitución en la Comunidad de Madrid, existe una asociación estadísticamente significativa entre la Edad (categorizada) y la Infección por VIH; es decir, la proporción de mujeres infectadas por VIH varía en función de la Edad (<25, 25-29, ≥30 años).

A continuación, calculamos el test estadístico. Para ello, debemos calcular en primer lugar las Frecuencias Esperadas en cada una de las celdas, utilizando la siguiente fórmula:

$$Frecuencia\ esperada = \frac{Total\ Fila \times Total\ Columna}{Total}$$

En la siguiente Tabla, se muestran las frecuencias observadas y las frecuencias esperadas en cada una de las celdas de la Tabla de contingencia:

		Infección por VIH		
		No	Si	Total
Edad (categorizada)	<25	137 145.6	17 8.4	154
	25-29	122 119.1	4 6.9	126
	≥30	157 151.3	3 8.7	160
Total		416	24	440

A continuación, calculamos el valor del test estadístico mediante la siguiente fórmula:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(137-145.6)^2}{145.6} + \frac{(17-8.4)^2}{8.4} + \frac{(122-119.1)^2}{119.1} + \frac{(4-6.9)^2}{6.9} + \\ + \frac{(157-151.3)^2}{151.3} + \frac{(3-8.7)^2}{8.7} = 14.56$$

Bajo la Hipótesis Nula, el estadístico chi-cuadrado sigue una distribución chi-cuadrado con (r-1)x(c-1) grados de libertad. r es el número de categorías de la variable representad en filas (variable de exposición) y c es el número de categorías de la variable representada en columnas (variable de interés).

En nuestro ejemplo, el cálculo de los grados de libertad de la distribución chi-cuadrado se haría como:

$$grados\ libertad = (r-1)\times(c-1) = (3-1)\times(2-1) = 2$$

El p-valor se calcularía como la probabilidad de observar una diferencia entre las frecuencias esperadas y observadas como la obtenida o más extrema, si la Hipótesis Nula fuera cierta; es decir,

$$p-valor = \Pr(\chi_2^2 \geq 14.56) < 0.001$$

El p-valor del test chi-cuadrado de asociación es <0.001; esto es, existe una asociación estadísticamente significativa entre la edad (categorizada) y la infección por VIH; es decir, la proporción de mujeres infectadas por VIH varía en función de la Edad.

5.3. Magnitud de la asociación para Tablas rx2

Una vez que hemos detectado la existencia de una asociación entre la Edad categorizada (<25, 25-29, ≥30 años) y la Infección por VIH, el siguiente paso consiste en cuantificar la magnitud de la asociación. Al igual que en el caso en que tanto la variable de exposición como la variable de interés son dicotómicas, hay 3 medidas para cuantificar la magnitud de la asociación entre una

variable de exposición con más de 2 categorías y una variable de interés dicotómica:

- Diferencia de proporciones (DP)
- Razón de proporciones (RP)
- Odds ratio (OR)

En el caso de variable de exposición y de interés dicotómica, hemos calculado la medida de asociación correspondiente comparando la proporción (o, la odds) de individuos que presentan la característica de interés en los individuos expuestos con la proporción (o, la odds) de individuos que presentan la característica de interés en los individuos no expuestos.

En el caso en el que la variable de exposición tiene más de 2 categorías, seleccionamos una categoría como la categoría de referencia, y comparamos la proporción (o, la odds) de individuos que presentan la característica de interés de cada categoría de la variable de exposición con la proporción (o, la odds) de individuos que presentan la característica de interés en la categoría de referencia.

El Test chi-cuadrado de asociación para Tablas rx2 permite determinar si existe una asociación estadísticamente significativa entre una variable de exposición con más de 2 categorías y una variable de interés dicotómica

Hay 3 medidas para cuantificar la magnitud de la asociación entre una variable de exposición con más de 2 categorías y una variable de interés dicotómica:

- Diferencia de proporciones (DP)
- Razón de proporciones (RP)
- Odds ratio (OR)

Diferencia de Proporciones

Supongamos que en nuestro ejemplo decidimos considerar como categoría de referencia a las mujeres menores de 25 años. En primer lugar, comparamos la proporción de mujeres infectadas por VIH entre las mujeres que tienen 25-29 años con la proporción de mujeres infectadas por VIH entre las mujeres menores de 25 años.

Para facilitar el cálculo, seleccionamos las filas de la Tabla de contingencia correspondientes a las categorías de Edad <25 y 25-29 años:

		Infección por VIH		
		No	Si	Total
Edad (categorizada)	<25	137	17	154
	25-29	122	4	126
Total		416	24	440

Si consideramos como No expuestas a las mujeres <25 años (categoría de referencia) y como Expuestas a las mujeres entre

En el caso en el que la variable de exposición tiene más de 2 categorías, seleccionamos una categoría como la categoría de referencia, y comparamos la proporción (o, la odds) de individuos que presentan la característica de interés de cada categoría de la variable de exposición con la proporción (o, la odds) de individuos que presentan la característica de interés en la categoría de referencia.

25 y 29 años, la diferencia de proporciones se calcularía como:

Proporción de mujeres infectadas por VIH entre las mujeres <25 años (p_0):

$$p_0 = d_0 / n_0 = 17/154 = 0.11$$

Proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años (p_1):

$$p_1 = d_1 / n_1 = 4/126 = 0.03$$

$$Diferencia\ \Pr oporciones\ (DP)_{25-29\ vs\ <25} = p_1 - p_0 = 0.03 - 0.11 = -0.08$$

En las mujeres de la muestra, la diferencia en la proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años y las <25 años es -0.08.

Si siguiéramos el procedimiento presentado en el apartado 4.3. para calcular un Intervalo de Confianza al 95% para la Diferencia en la Proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años y las mujeres <25 años, obtendríamos:

$$IC_{95\%}(DP)_{25-29\ vs\ <25} = (-0.14;\ -0.02)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la diferencia en la proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años y las mujeres <25 años está entre -0.14 y -0.02.

Como el Intervalo de Confianza para la diferencia de proporciones no incluye al 0, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años NO es la misma que la proporción de mujeres infectadas por VIH entre las mujeres <25 años.

En segundo lugar, comparamos la proporción de mujeres infectadas por VIH entre las mujeres que tienen ≥30 años con la proporción de mujeres infectadas por VIH entre las mujeres menores de 25 años.

Para facilitar el cálculo, seleccionamos las filas de la Tabla de contingencia correspondientes a las categorías de Edad <25 y ≥30 años:

		Infección por VIH		
		No	Si	Total
Edad (categorizada)	<25	137	17	154
	≥30	157	3	160

Si consideramos como No expuestas a las mujeres <25 años (categoría de referencia) y como Expuestas a las mujeres de 30 años ó más, la diferencia de proporciones se calcularía como:

Proporción de mujeres infectadas por VIH entre las mujeres <25 años (p_0):

$$p_0 = d_0 / n_0 = 17/154 = 0.11$$

Proporción de mujeres infectadas por VIH entre las mujeres ≥30 años (p_2):

$$p_2 = d_2 / n_2 = 3/157 = 0.02$$

$$Diferencia\ \Pr oporciones\ (DP)_{\geq 30\ vs\ <25} = p_2 - p_0 = 0.02 - 0.11 = -0.09$$

En las mujeres de la muestra, la diferencia en la proporción de mujeres infectadas por VIH entre las mujeres ≥30 años y las <25 años es -0.09.

Un Intervalo de Confianza al 95% para la Diferencia en la Proporción de mujeres infectadas por VIH entre las mujeres ≥30 años y las mujeres <25 años sería:

$$IC_{95\%}(DP)_{\geq 30\ vs\ <25} = (-0.15; -0.04)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la diferencia en la proporción de mujeres infectadas por VIH entre las mujeres ≥30 años y las mujeres <25 años está entre -0.15 y -0.04.

Como el Intervalo de Confianza para la diferencia de proporciones no incluye al 0, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la proporción de mujeres infectadas por VIH entre las mujeres ≥30 años NO es la misma que la proporción de mujeres infectadas por VIH entre las mujeres <25 años.

Razón de Proporciones

Siguiendo el procedimiento presentado previamente, la razón entre la proporción de mujeres infectadas por VIH en las mujeres entre 25 y 29 años y la proporción de mujeres infectadas por VIH en las mujeres menores de 25 años se calcularía como:

$$Razón\ \text{Pr}oporciones(RP)_{25-29\ vs.\ <25} = p_1 / p_0 = 0.03 / 0.11 = 0.27$$

En la muestra, la proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años es un 73% (1-0.27) menor que entre las mujeres <25 años.

$$IC_{95\%}(RP)_{25-29\ vs\ <25} = (0.10;\ 0.83)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años es entre un 17% (1-0.83) y un 90% (1-0.10) menor que entre las mujeres <25 años.

Como el Intervalo de Confianza para la razón de proporciones no incluye al 1, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la proporción de mujeres infectadas por VIH entre las mujeres de 25-29 años NO es la misma que la proporción de mujeres infectadas por VIH entre las mujeres <25 años.

Del mismo modo, la razón entre la proporción de mujeres infectadas por VIH en las mujeres de 30 años ó más y la proporción de mujeres infectadas por VIH en las mujeres menores de 25 años se calcularía como:

$$Razón\ \text{Pr}oporciones(RP)_{\geq 30\ vs.\ <25} = p_2 / p_0 = 0.02 / 0.11 = 0.18$$

En la muestra, la proporción de mujeres infectadas por VIH entre las mujeres ≥30 años es un 82% (1-0.18) menor que entre las mujeres <25 años.

$$IC_{95\%}(RP)_{\geq 30 \ vs \ <25} = (0.05;\ 0.57)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la proporción de mujeres infectadas por VIH entre las mujeres ≥30 años es entre un 43% (1-0.57) y un 95% (1-0.05) menor que entre las mujeres <25 años.
Como el Intervalo de Confianza para la razón de proporciones no incluye al 1, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la proporción de mujeres infectadas por VIH entre las mujeres ≥30 años NO es la misma que la proporción de mujeres infectadas por VIH entre las mujeres <25 años.

Odds ratio

La odds ratio comparando la odds de infección por VIH en las mujeres entre 25 y 29 años con la odds de infección por VIH en las mujeres menores de 25 años se calcularía como:

$$Odds\ Ratio\ (OR)_{25-29\ vs.\ <25} = \frac{odds_{25-29}}{odds_{<25}} = \frac{d_1/h_1}{d_0/h_0} = \frac{4/122}{17/137} = 0.26$$

En la muestra, la odds de infección por VIH en las mujeres de 25-29 años es un 74% (1-0.26) menor que entre las mujeres <25 años; las mujeres de 25-29 años tienen un "riesgo" de infección por VIH un 74% menor que las mujeres <25 años.

$$IC_{95\%}(OR)_{25-29\ vs\ <25} = (0.09;\ 0.82)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la odds de infección por VIH en las mujeres de 25-29 años es entre un 18% (1-0.82) y un 91% (1-0.09) menor que entre las mujeres <25 años.

Como el Intervalo de Confianza para la OR no incluye al 1,

estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la odds (y, por tanto, la proporción) de mujeres infectadas por VIH entre las mujeres de 25-29 años NO es la misma que la odds (y, por tanto, la proporción) de mujeres infectadas por VIH entre las mujeres <25 años.

La odds ratio comparando la odds de infección por VIH en las mujeres de 30 ó más años con la odds de infección por VIH en las mujeres menores de 25 años se calcularía como:

$$Odds\ Ratio\ (OR)_{\geq 30\ vs.\ <25} = \frac{odds_{\geq 30}}{odds_{<25}} = \frac{d_2 / h_2}{d_0 / h_0} = \frac{3/157}{17/137} = 0.15$$

La asociación entre dos variables categóricas cualquiera puede explorarse de forma descriptiva mediante una Tabla de contingencia, y a continuación puede realizarse el test chi-cuadrado de asociación para determinar si en la población de la que se extrajeron las muestras, existe una asociación estadísticamente significativa entre ambas variables.

En la muestra, la odds de infección por VIH en las mujeres ≥30 años es un 85% (1-0.15) menor que entre las mujeres <25 años; las mujeres ≥30 años tienen un "riesgo" de infección por VIH un 85% menor que las mujeres <25 años.

$$IC_{95\%}(OR)_{\geq 30\ vs\ <25} = (0.04;\ 0.55)$$

Estamos seguros al 95% de que en la población de mujeres que se dedican a la prostitución en la comunidad de Madrid, la odds de infección por VIH en las mujeres ≥30 años es entre un 45% (1-0.55) y un 96% (1-0.04) menor que entre las mujeres <25 años.

Como el Intervalo de Confianza para la OR no incluye al 1, estamos seguros al 95% de que en la población de la que se extrajeron las muestras, la odds (y, por tanto, la proporción) de mujeres infectadas por VIH entre las mujeres ≥30 años NO es la misma que la odds (y, por tanto, la proporción) de mujeres infectadas por VIH entre las mujeres <25 años.

6. Asociación entre dos variables categóricas

En algunas situaciones, podemos estar interesados en estudiar la posible asociación entre una variable de exposición con r categorías y una variable de interés con c categorías. Por ejemplo, supongamos que ha diseñado un estudio en pacientes con tumores cerebrales con el objetivo de estudiar la posible asociación entre la localización del tumor (lóbulo frontal, lóbulo temporal, otras áreas) y la naturaleza del mismo (benigno, maligno, otros).

Los métodos presentados previamente pueden extenderse para el estudio de la asociación entre dos variables categóricas cualquiera. En primer lugar, construiríamos una Tabla de contingencia y a continuación realizaríamos el test chi cuadrado para determinar si en la población de la que se extrajeron las muestras, existe una asociación estadísticamente significativa entre ambas variables.

El único problema que surge en el caso en el que la variable de interés tiene más de dos categorías, es que no es posible cuantificar la magnitud de la asociación entre la variable de exposición y la variable de interés. Por eso, es recomendable, intentar trabajar con variables de interés dicotómicas; si la variable de interés tiene más de dos categorías, podemos categorizarla de forma que tenga únicamente 2 categorías, y aplicar los procedimientos presentados previamente.

El único problema que surge en el caso en el que la variable de interés tiene más de dos categorías, es que no es posible cuantificar la magnitud de la asociación entre la variable de exposición y la variable de interés.

Conclusiones

En este tema se han descrito los métodos utilizados para el análisis de variables de interés dicotómicas. Específicamente, se han presentado los métodos para (1) estimar la proporción de individuos que presentan una determinada característica de interés en la población a estudio, (2) determinar qué variables de exposición se asocian con la variable de interés dicotómica, y (3) cuantificar la magnitud de la asociación entre la variable de exposición y la variable de interés dicotómica, en la población a estudio.

Referencias bibliográficas

1. *Peña D, Romo J. Introducción a La Estadística para las Ciencias Sociales. Editorial McGraw Hill, 2003*

2. *Martínez M.Bioestadística amigable. Editorial Díaz de Santos, 2006*

3. *Hernández-Aguado I, Gil A, Delgado M, Bolumar F. Manual de Epidemiología y Salud Pública. Editorial Médica Panamericana, 2005*

4. *Kirkwood B, Sterne J. Essential Medical Statistics. Blackwell Science Ltd, 2001.*