

## 6A. RELACIÓN ENTRE DOS VARIABLES CATEGÓRICAS.

### Formato de las tablas de contingencia

La descripción y comparación de la distribución de frecuencias de una variable categórica para cada una de las categorías de otra variable categórica, se realiza mediante una tabla de doble entrada, llamada *Tabla de contingencia*, en cuyas casillas se muestra su distribución conjunta. Las variables exposición y respuesta se pueden asignar tanto a filas como a columnas y colocar la situada en filas en orden ascendente o descendente. También se pueden pedir porcentajes de filas, columnas o totales. El resultado mantiene su validez siempre y cuando la tabla se lea e interprete de forma correcta. Para facilitar su correcta lectura e interpretación es preferible utilizar una sistemática uniforme en las tablas de contingencia. En el curso se utilizarán de la forma que se muestra en la Tabla 6A.1. La variable exposición (o de segmentación) se colocará en filas y en orden descendente (de código mayor a menor). La variable respuesta (o a describir) se sitúa en columnas. Para continuar con la misma “lógica” se debe especificar porcentajes observados de fila, excepto en diseños caso-control, en el que los porcentajes por fila no tienen sentido y sí los porcentajes por columna (nos darían los % de exposición y no exposición en los casos y en los controles). Los valores totales de fila y de columna se denominan marginales.

	Respuesta No	Respuesta Sí	Total
Exposición Sí	<b>b</b>	<b>a</b>	<b>a+b</b>
Exposición No	<b>d</b>	<b>c</b>	<b>c+d</b>
Total	<b>b+d</b>	<b>a+c</b>	<b>a+b+c+d</b>

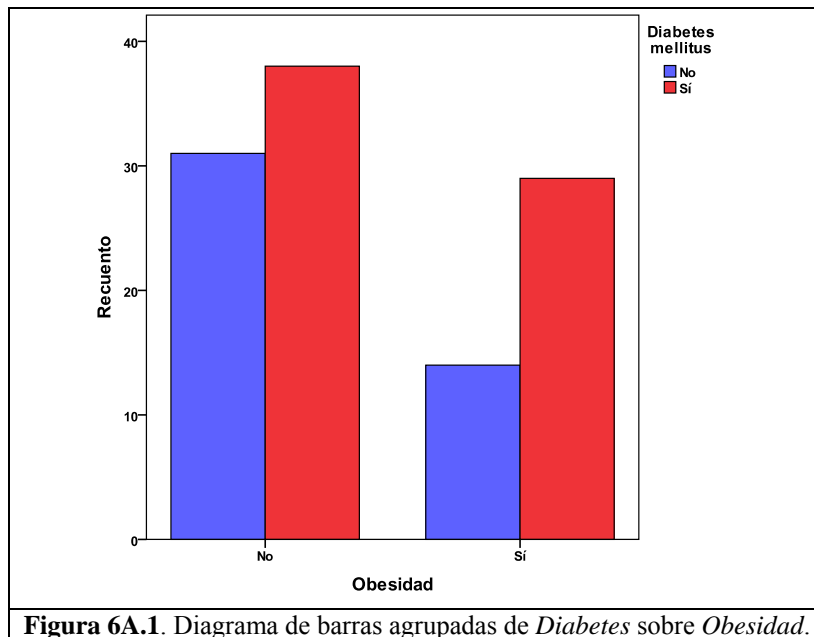
### Relación entre exposición binaria - respuesta binaria: comparación de dos proporciones

El análisis de la relación entre una variable respuesta binaria como *Diabetes Sí/No* con una exposición binaria como *Obesidad Sí/No* consiste en la *comparación de dos proporciones*: la proporción de diabetes en los obesos frente a la proporción de diabetes en los no obesos, según puede apreciarse en las Tablas 6A.2 y 6A.3. Si la proporción de diabetes en los obesos es distinta que en los no obesos, las dos variables están relacionadas o son estadísticamente dependientes. Por el contrario, si la proporción de diabetes es similar en obesos y en no obesos, las dos variables no están relacionadas o son estadísticamente independientes. Según se aprecia en la Tabla 6A.2, en los obesos hay mayor proporción de diabéticos (67,4%) que en los no obesos (55,1%), lo que sugiere existencia de asociación.

			Diabetes		Total
			No	Sí	
Obesidad	Sí	n % Sí	<b>14</b> 32,6%	<b>29</b> <b>67,4%</b>	43 100%
	No	n % No	<b>31</b> 44,9%	<b>38</b> <b>55,1%</b>	69 100%
	Total	n % Total	45 40,2%	67 59,8%	112 100%

Esta relación se puede representar gráficamente con un *Diagrama de barras agrupado* (Figura 6A.1), con la distribución de la variable respuesta (*Diabetes*) en cada categoría de la variable exposición (*Obesidad*). Se puede apreciar que la proporción de diabetes en obesos es mayor que en no obesos, aunque por los recuentos absolutos puede dar la impresión de lo contrario. Las pruebas estadísticas utilizadas en el análisis de esta relación y sus condiciones de aplicación, se muestran en la Tabla 6A.3.

<b>Respuesta binaria – Exposición binaria</b> <i>Diabetes Sí/No – Obesidad Sí/no</i>	
<b>Comparación de 2 proporciones</b> <i>Proporción de Diabetes en Obesos</i> <i>Proporción de Diabetes en No obesos</i>	
<b>Número de casillas con frecuencia esperada menor de 5</b>	
<b>Menos del 20% (Ninguna)</b>	<b>Más del 20% (Alguna)</b>
<b>Prueba de chi cuadrado</b> <b>Razón de verosimilitud</b> <b>Corrección de continuidad de Yates</b>	<b>Prueba exacta de Fisher bilateral</b>



**Figura 6A.1.** Diagrama de barras agrupadas de *Diabetes* sobre *Obesidad*.

Se debe evaluar cuantas casillas tienen una frecuencia esperada menor de 5. La frecuencia esperada es la que tendría la tabla en caso de no haber diferencias (50% de diabéticos en obesos y en no obesos). Si el número de casillas con frecuencia esperada menor de 5 es menor del 20% (ninguna casilla en el caso de una tabla de 2x2) se pueden utilizar la prueba de chi-cuadrado, la razón de verosimilitud y la corrección de continuidad de Yates. En este caso la frecuencia mínima esperada es de 17,28. Por tanto se puede utilizar estas tres pruebas.

- **Prueba de chi-cuadrado** ( $\chi^2$ ) de Pearson. La ley de probabilidad de chi-cuadrado ofrece la probabilidad de ocurrencia de todas las tablas de contingencia de muestras de 112 casos y con marginales fijos (43 y 69 en filas y 45 y 67 en columnas en este caso) obtenidas de una población sin relación entre las variables (hipótesis nula: 50% de diabetes en obesos y en no obesos). El resultado de la prueba de chi-cuadrado en esta muestra concreta es no significativo: Chi-cuadrado=1,686; gl=1; p = 0,194. La probabilidad de obtener una diferencia de proporciones igual o mayor a la obtenida en esta muestra (67,4 – 55,1 = 12,3%) en el supuesto de que no haya diferencias en la población origen es de 0,194. Por tanto no hay diferencias estadísticamente significativas en la proporción de diabetes en obesos y en no obesos.
- **Razón de verosimilitudes.** También se puede emplear esta prueba clásica en el análisis de datos categóricos utilizada en la regresión logística. Su resultado es muy similar al de la chi-cuadrado, siendo casi idéntico en muestras grandes. En este caso la prueba de la razón de verosimilitudes es p = 0,272.
- **Corrección de continuidad** de Yates. Consiste en una modificación conservadora en la prueba de chi-cuadrado que actualmente no tiene aplicación, aunque todavía se puede encontrar algún revisor que la solicite, sobre todo en tamaños de muestra menor de 200 que es donde pudiera tener alguna aplicación. En este ejemplo la prueba de chi-cuadrado con la corrección de continuidad es p = 0,465.
- **Prueba exacta de Fisher** bilateral. En caso de que más del 20% de las casillas (alguna de las casillas en una tabla de 2x2) tengan una frecuencia esperada mayor del 20% no se pueden aplicar las prueba de chi-cuadrado, la razón de verosimilitud ni la corrección de continuidad, y debe utilizarse la *prueba exacta de Fisher* bilateral, cuyo resultado en este caso es p = 0,466.

Las pruebas estadísticas nos informan del grado en que la relación puede ser debida al azar, pero no informa sobre la magnitud del efecto o la importancia clínica o práctica. La magnitud del efecto de la relación entre dos variables binarias se valora con las medidas de frecuencia (proporción, odds), de asociación (riesgo relativo, odds ratio) e impacto (diferencia de riesgos, NNT) descritas en la unidad 8.

## Relación entre exposición politómica - respuesta binaria: comparación de varias proporciones

El análisis de la relación entre una variable respuesta binaria como *Diabetes Sí/No* con una predictora categórica con más de dos categorías (politómica) como *ObesidadOr* (*Nivel de obesidad Imc normal/Sobrepeso/Obesidad*) consiste en la *comparación de varias proporciones*: la proporción de diabéticos en los que tienen Imc Normal, en los que tienen Sobrepeso y en los que tienen Obesidad (Tablas 6A.4 y 6A.5). Las proporciones de diabetes en Imc Normal (38,9%), Sobrepeso (60,8%) y Obesidad (67,4%) mostradas en la tabla de contingencia 3x2 de la Tabla 6A.4 y el gráfico de barras agrupadas de la Figura 6A.2 sugieren la existencia de asociación.

		Diabetes		Total	
		No	Sí		
<b>ObesidadOr</b>	Obesidad	n % Obesidad	14 32,6%	29 <b>67,4%</b>	43 100%
	Sobrepeso	n % Sobrepeso	20 39,2%	31 <b>60,8%</b>	51 100%
	Imc Normal	n % Imc Normal	11 61,1%	7 <b>38,9%</b>	18 100%
Total		n % Total	45 40,2%	67 59,8%	112 100%

Las pruebas estadísticas utilizadas en el análisis de esta relación y sus condiciones de aplicación se muestran en la Tabla 6A.5. Se debe evaluar cuantas casillas tienen una frecuencia esperada menor de 5. La frecuencia esperada es la que tendría la tabla en caso de no haber diferencias (igual proporción de diabéticos en las tres categorías de ObesidadOr). Si el número de casillas con frecuencia esperada menor de 5 es menor del 20% (una o ninguna casilla en el caso de una tabla de 3x2) se pueden utilizar la prueba de chi-cuadrado y la razón de verosimilitud. En este caso la frecuencia mínima esperada es de 7,23. Por tanto se puede utilizar estas dos pruebas.

<b>Respuesta binaria – Exposición politómica</b> <i>Diabetes Sí/No – Nivel Obesidad Imc normal/Sobrepeso/Obesidad</i>	
<b>Comparación de 5 proporciones</b> <i>Proporción de Diabetes en Imc normal</i> <i>Proporción de Diabetes en Sobrepeso</i> <i>Proporción de Diabetes en Obesidad</i>	
<b>Número de casillas con frecuencia esperada menor de 5</b>	
<b>Menos del 20% (Una o ninguna)</b>	<b>Más del 20% (Dos o más)</b>
<b>Prueba de chi cuadrado</b> <b>Razón de verosimilitud</b> <b>Asociación lineal (si Ordinal)</b>	<b>Prueba exacta de Fisher bilateral</b> <b>(o Agrupación de categorías)</b>
<b>Si significativa: Comparaciones múltiples</b>	

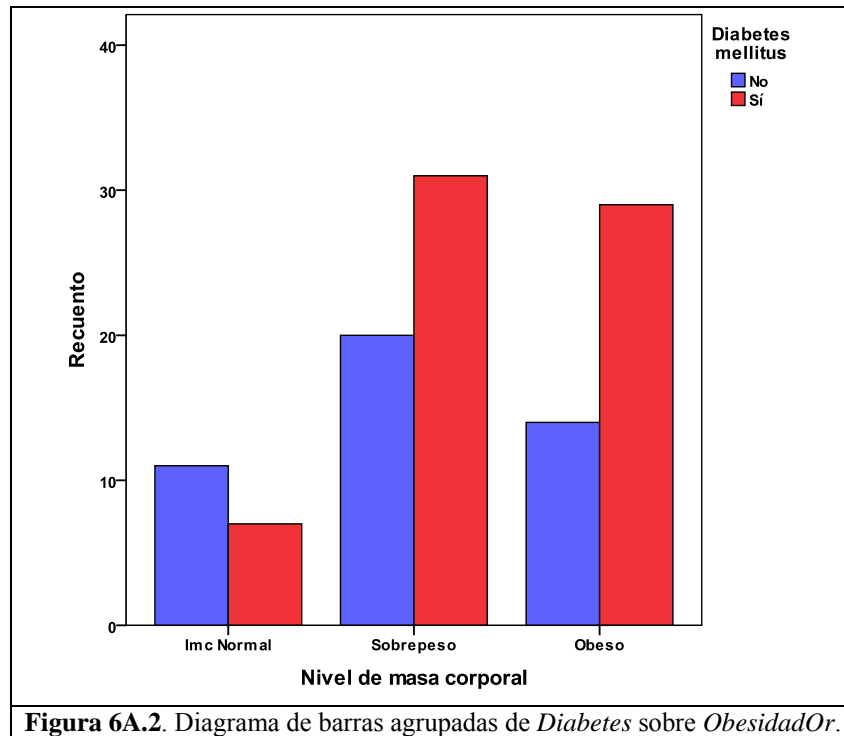
En este caso el resultado de la generalización de la *prueba de chi-cuadrado* a una tabla de contingencia de 3x2 es Chi-cuadrado=4,340; gl=2; p=0,114. Indica que no hay diferencias estadísticamente significativas en la proporción de diabéticos en las tres categorías de obesidad.

El resultado de la prueba de *Razón de verosimilitudes* es muy parecido a la chi-cuadrado: p=0,118. Ambas requieren que existan menos del 20% de casillas con frecuencias esperadas < 5 (en la tabla de 3x2 se permite una de las seis casillas con frecuencia esperada < 5). En este caso se cumple este requisito. Aquí no es aplicable la corrección de continuidad que es propia de una tabla de 2x2.

Por tratarse de una variable ordinal, se puede aplicar la prueba de tendencia de *Asociación lineal*. A diferencia de las pruebas anteriores que valoran la igualdad de las proporciones, esta prueba valora si hay una tendencia lineal creciente o decreciente de la proporción de diabetes entre las categorías ordenadas de ObesidadOr. Las tres proporciones de Diabetes sugieren una tendencia creciente: Imc normal < Sobrepeso < Obesidad. Requiere como condición de aplicación que todas las frecuencias esperadas sean  $\geq 2$ , siempre que las frecuencias cercanas a 2 no se presenten en casillas contiguas. En este caso la prueba de tendencia de asociación lineal es p=0,055. Indica que no existe una tendencia lineal estadísticamente significativa de las proporciones de diabetes en los tres niveles de obesidad, aunque por poco. Esta prueba sólo es aplicable en el caso de variables exposición con escala de medida ordinal. No tiene sentido en variables exposición con escala de medida nominal.

Cuando la frecuencia esperada es  $< 5$  en más del 20% de las casillas no se puede utilizar la prueba de chi-cuadrado. En estos casos se pueden utilizar **pruebas exactas** (como la prueba exacta de Fisher o la regresión logística exacta), no siempre disponibles en los programas estadísticos habituales, o bien **agrupar categorías** con frecuencias esperadas bajas hasta que sean superiores a 5 y se puedan aplicar las pruebas de chi-cuadrado. La agrupación de categorías se puede hacer por criterio clínico (la que tenga más sentido) o estadístico (la que agrupe las categorías de muestra más pequeña).

En caso de ser significativa, la conclusión de la comparación de estas tres proporciones es demasiado general: simplemente indica que algunas de las tres proporciones son estadísticamente diferentes. Se puede avanzar más en el análisis de esas diferencias realizando comparaciones múltiples, como se verá en la unidad 7.



### Videotutorial 6A1Contingencia1.avi

Se muestra como se analiza la relación entre la variable exposición binaria *Obesidad Sí/No* y la variable respuesta binaria *Diabetes Sí/No*. El cuadro *Tabla de contingencia* permite crear la tabla de contingencia 2x2, ofrece el gráfico de barras agrupadas y las pruebas estadísticas de chi-cuadrado, razón de verosimilitud, corrección de continuidad y exacta de Fisher unilateral y bilateral. También ofrece el número y porcentaje de casillas con frecuencia esperada menor de 5.

También se muestra como se analiza la relación entre la variable exposición ordinal *Nivel Obesidad Imc normal/Sobrepeso/Obesidad* y la variable respuesta binaria *Diabetes Sí/No* con el cuadro *Tabla de contingencia* y las pruebas estadísticas de chi-cuadrado (en la tabla de 3x2 ofrece las pruebas chi-cuadrado, razón de verosimilitud y asociación lineal). Si se dispone del módulo *Exactas*, se puede obtener la prueba exacta de Fisher unilateral y bilateral.