Unidad 6, Lección 3

Asociaciones en datos categóricos

Busquemos relaciones entre variables categóricas.

3.1 Inglés o Matemáticas

La tabla muestra el curso preferido y la mano dominante (si la persona es zurda o diestra) de una muestra de 300 personas.

	prefiere Inglés	prefiere Matemáticas	total
es zurda	10	20	30
es diestra	90	180	270
total	100	200	300

Para cada uno de estos cálculos, interpreta el porcentaje describiéndolo en términos de la situación.

1. 10%, que se obtiene al calcular: $\frac{10}{100} = 0.1$

2. 67%, que se obtiene al calcular: $\frac{180}{270} \approx 0.67$

3. 30%, que se obtiene al calcular: $\frac{90}{300} = 0.3$

3.2 Asociaciones en datos categóricos

1. La tabla de doble entrada muestra datos de 55 lugares distintos que tienen corales. Los científicos tienen una lista de los posibles químicos que podrían afectar la salud del coral. Comienzan por estudiar cómo podría estar relacionada la concentración de nitrato con el estado de salud del coral. La tabla muestra el estado de salud del coral (saludable o enfermo) y la concentración de nitrato (baja o alta).

	concentración de nitrato baja	concentración de nitrato alta	total
saludable	20	5	25
enfermo	8	22	30
total	28	27	55

   a. Completa esta tabla de doble entrada de frecuencias relativas usando los datos de la tabla de doble entrada. Las frecuencias relativas se calculan usando el total de cada columna.

	concentración de nitrato baja	concentración de nitrato alta
saludable		
enfermo		
total	100%	100%

   b. Cuando hay una concentración de nitrato baja, ¿cuál tiene una frecuencia relativa más alta: el coral saludable o el enfermo?

   c. Cuando hay una concentración de nitrato alta, ¿cuál tiene una frecuencia relativa más alta: el coral saludable o el enfermo?

d. Según estos datos, ¿hay una posible **asociación** entre el estado de salud del coral y el nivel de concentración de nitrato? Explica tu razonamiento.

e. Luego, los científicos estudian cómo podría estar relacionada la concentración de dióxido de silicio con el estado de salud del coral. En la tabla se muestran las frecuencias relativas que se calcularon usando el total de cada columna. Según estos datos, ¿hay una posible asociación entre el estado de salud del coral y el nivel de concentración de dióxido de silicio? Explica tu razonamiento.

	concentración de dióxido de silicio baja	**concentración de dióxido de silicio alta**
saludable	44%	46%
enfermo	56%	54%
total	100%	100%

2. Jada encuestó a 300 personas de distintos grupos de edad acerca de sus zapatos preferidos. La tabla de doble entrada resume los resultados de la encuesta.

	prefiere zapatos deportivos sin cordones	**prefiere zapatos deportivos con cordones**	**prefiere zapatos que no sean deportivos**	**total**
4 a 10 años	21	12	3	36
11 a 17 años	21	48	39	108
18 a 24 años	15	54	87	156
total	57	114	129	300

Jada concluye que hay una posible asociación entre la edad y los zapatos preferidos. ¿Es razonable la conclusión de Jada? Explica tu razonamiento.

3. La tabla de doble entrada resume los datos sobre las herramientas de escritura preferidas y la mano dominante de una muestra de 100 personas.

	es zurda	es diestra	total
prefiere el bolígrafo	7	82	89
prefiere el lápiz	6	5	11
total	13	87	100

¿Hay una posible asociación entre la mano dominante y la herramienta de escritura preferida? Explica tu razonamiento.

¿Estás listo para más?

La tabla de doble entrada, que está incompleta, muestra los resultados de una encuesta acerca del tipo de tratamiento de medicina deportiva y el tiempo de recuperación de 33 atletas estudiantiles que visitaron al entrenador deportivo.

	volvió a jugar en menos de 2 días	**volvió a jugar en 2 días o más**
se trató con hielo	8	4
se trató con calor		

1. Elige 2 valores para completar la tabla de doble entrada que permitan concluir que hay una asociación entre volver a jugar en menos de 2 días y el tratamiento (hielo o calor). Explica tu razonamiento.

2. Elige 2 valores para completar la tabla de doble entrada que permitan concluir que no hay una asociación entre volver a jugar en menos de 2 días y el tratamiento (hielo o calor). Explica tu razonamiento.

3. ¿Cuáles valores fueron más fáciles de escoger: los 2 valores que muestran que hay una asociación o los 2 valores que muestran que no hay una asociación? Explica tu razonamiento.

3.3 Asocien sus propias variables

1. En grupo, identifiquen un par de variables categóricas que piensen que podrían estar asociadas y otro par que piensen que no estarían asociadas.

2. Imaginen que su grupo recolectó datos para cada par de variables categóricas. Creen una tabla de doble entrada que pueda representar cada conjunto de datos. Para completar cada tabla, inventen datos que tengan un valor total de 100. Recuerden que una tabla muestra que hay una posible asociación y la otra tabla muestra que no hay asociación.

3. Expliquen o muestren por qué parece que hay una asociación entre el primer par de variables y por qué parece que no hay una asociación entre el segundo par de variables.

4. Preparen una presentación de su trabajo para compartirla.

Resumen de la lección 3

Decir que hay una **asociación** entre dos variables significa que las dos variables se relacionan estadísticamente entre sí. Por ejemplo, podríamos esperar que las ventas de helados fueran mayores en los días soleados que en los días de nieve. Si hubiera mayores ventas en los días soleados que en los días de nieve, entonces diríamos que hay una posible asociación entre las ventas de helados y el hecho de que haga sol o caiga nieve. Cuando se trata con variables categóricas, a menudo se usan tablas de frecuencias relativas por fila o por columna para buscar asociaciones en los datos.

Esta tabla de doble entrada muestra el estado del tiempo y las ventas de conos en una heladería en particular a lo largo de 41 días.

	día soleado	día de nieve	total
se vendieron menos de 50 conos	8	7	15
se vendieron 50 conos o más	22	4	26
total	30	11	41

Puede ser difícil identificar un patrón en los datos sin procesar. En especial, cuando los totales de la fila o de la columna no son los mismos para categorías diferentes. Por esto, la tabla de datos anterior se debe convertir en una tabla de frecuencias relativas por fila o por columna para comparar mejor las categorías. Observa que en esta heladería, el número de días con ventas bajas es aproximadamente el mismo para los dos tipos de clima, lo que contradice nuestra intuición. En este caso, tiene sentido examinar el porcentaje de días con ventas altas para cada estado del tiempo por separado. Es decir, considerar las frecuencias relativas por columna.

	día soleado	día de nieve
se vendieron menos de 50 conos	27%	64%
se vendieron 50 conos o más	73%	36%
total	100%	100%

A partir de la tabla de frecuencias relativas por columna, queda claro que en la mayoría de los días soleados (73%) se vendieron al menos 50 conos, mientras que en la mayoría de los días de nieve (64%) se vendieron menos de 50 conos. Estos porcentajes son muy diferentes, lo que sugiere que hay una asociación entre el clima y el número de conos vendidos. Una panadería podría preguntarse si el clima afecta también sus ventas de *muffins*.

	día soleado	**día de nieve**
se vendieron menos de 50 *muffins*	32%	35%
se vendieron 50 *muffins* o más	68%	65%
total	100%	100%

En el caso de la panadería, parece que no hay una asociación entre el estado del tiempo y las ventas de *muffins*, ya que los porcentajes de días con ventas bajas son muy similares para los distintos estados del tiempo y los porcentajes son muy cercanos en los días que venden muchos *muffins*.

Usar tablas de frecuencias relativas por fila o por columna ayuda a organizar los datos para poder comparar fácilmente las columnas (o filas) que representan las distintas categorías de una variable. Esta comparación se puede hacer usando una tabla de doble entrada, pero es necesario tener en cuenta las diferencias en el número de valores de cada categoría.