ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Septiembre 2022

ORIGINAL

Paso a paso. Prueba de la ji-cuadrado para independencia.

Molina M.

Hospital Infantil Universitario La Paz.

Resumen

La prueba de la ji-cuadrado para independencia compara las proporciones de dos variables cualitativas para determinar si son o no independientes.

Introducción

La prueba de la ji-cuadrado para independencia compara las proporciones de dos variables cualitativas para determinar si son o no independientes.

¿Quién no conoce la historia del Titanic, el barco insumergible que, finalmente, se sumergió en las gélidas aguas del Atlántico Norte?

Cuando se construyó, el RMS Titanic era el mayor barco de pasajeros de la época. Además de todos sus lujos y comodidades, contaba con las medidas de seguridad más avanzadas de su tiempo, como los mamparos del casco y las compuertas estancas, que debían asegurar la imposibilidad de que el buque pudiese hundirse.

Sin embargo, tras zarpar de Southampton el 10 de abril de 1912, el 14 de abril a las 23:40 horas chocó con un iceberg a 600 kilómetros al sur de Terranova, hundiéndose 3 horas después y causando la muerte de 1496 de las 2208 personas que viajaban a bordo.

La historia del Titanic nos enseña, entre otras muchas cosas, lo malo que es el exceso de confianza. Nadie podía imaginar que un barco de la época pudiese hundirse de esa manera. Probablemente fue ese exceso de confianza el causante de que no se valorasen adecuadamente los peligros del viaje, hasta que fue demasiado tarde.

Pero también nos enseña de qué forma tan enorme puede variar la valoración de un suceso según el punto de vista desde el que se haga. Lo que, para los pasajeros y para toda la Humanidad, fue una tragedia inmensa, para las langostas de la pecera del Titanic fue un auténtico milagro.

Un naufragio con clase

Una de las razones del alto número de muertos en el naufragio del Titanic lo causó el hecho de que el barco no tuviese botes salvavidas ni para la mitad de los pasajeros. Aunque los botes eran los más innovadores de la época, la cantidad se determinaba, de manera

incomprensible, por el peso del barco y no por el número de viajeros.

Además, siempre ha existido cierta controversia sobre si pobres y ricos corrieron la misma suerte ante la desgracia. Aunque, como suele decirse, todos iban en el mismo barco, y hay cosas que el dinero no puede comprar, cuando uno analiza las cifras de supervivientes no puede evitar sentir cierta inquietud.

Como es lógico, el porcentaje de fallecidos en las tres clases no es exactamente el mismo, pero ¿pueden las diferencias observadas deberse al azar o realmente influyó la clase en la probabilidad de supervivencia?

Vamos a tratar de responder a esta pregunta utilizando para ello el programa R, con su interfaz R-Commander y un conjunto de datos denominado TitanicSurvival.

Unos preparativos previos

Como ya sabréis, R funciona a base de paquetes, que son una especie de librerías con determinadas funcionalidades.

Muchos de estos paquetes incluyen conjuntos de datos (datasets), que pueden utilizarse para ensayar las técnicas de estadística que deseemos. Vamos a utilizar el dataset TitanicSurvival, del paquete carData. Este conjunto de datos incluye información de 1309 pasajeros, recogiendo en cuatro campos si sobrevivieron o no, su sexo, la edad y la clase en la que viajaban.

Primero, lanzamos R. Segundo, lanzamos R-Commander con el comando library (Rcmdr).

En la figura 1 os indico como cargar los datos una vez abierta la interfaz de R-Commander. Vamos al menú Datos->Cargar datos en paquetes->Leer conjunto de datos desde paquete adjunto… En la ventana emergente, seleccionáis carData como paquete y TitanicSurvival como conjunto de Datos. Pulsáis Aceptar y ya tendréis los datos en el conjunto de datos activo.

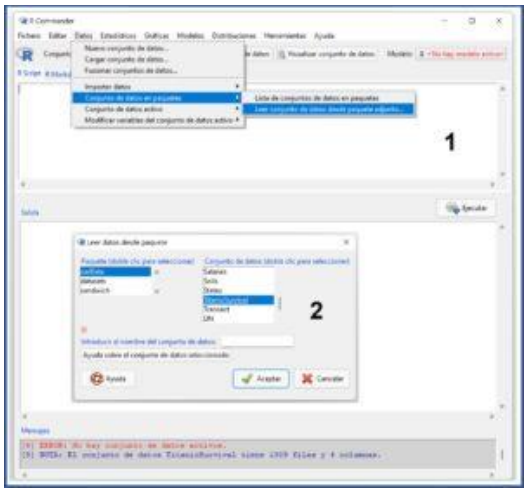

Figura 1. Carga del conjunto de datos.

Si no sabéis a qué paquete pertenece el conjunto de datos o si queréis ver qué datasets vienen cargados en R, podéis ver la lista completa seleccionando la opción del menú Datos->Cargar datos en paquetes->Lista de conjuntos de datos en paquetes.

Paso 1. Análisis descriptivo de los datos

En primer lugar, vamos a representar gráficamente las dos variables. Como se trata de dos variables nominales, realizaremos un diagrama de barras. Seleccionamos la opción del menú Gráficos->Gráfica de barras… (figura 2). En la ventana emergente marcamos la variable “survived” y pulsamos en el botón “Gráfica por grupos” para seleccionar “passengerClass”.

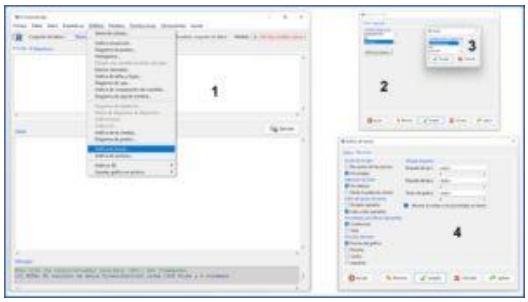

Figura 2. Elaboración del diagrama de barras.

Una vez hecho esto, abrimos la pestaña "Opciones" y marcamos las opciones "Escala de ejes" por porcentajes y "Estilo del grupo de barras" lado a lado. Pulsamos aceptar y obtenemos el diagrama de la figura 3.

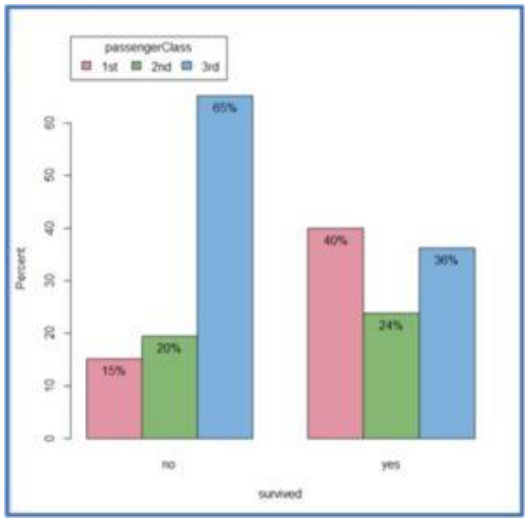

Figura 3. Diagrama de barras.

A simple vista, parece que la distribución de clases no es la misma entre los que sobreviven y los que no. Pasaremos a continuación a realizar nuestro contraste de hipótesis para ver si estas diferencias son estadísticamente significativas o pueden ser explicadas por puro azar.

Ya vimos en una entrada anterior que la elección de la prueba estadística depende del tipo de variables a comparar, de si tratamos con muestras independientes o con datos apareados y, en algunos casos, de la distribución de probabilidad que siguen los datos.

En este caso queremos comparar dos variables cualitativas: supervivencia (survived), con dos categorías (sí y no), y clase (passengerClass), con tres categorías (primera segunda y tercera).

Además, queremos saber si estas dos variables son independientes entre sí o si están relacionadas, de forma que el valor de una de ellas influya en el valor de la otra. Para hacer esto, las dos elecciones más sencillas son la prueba de la ji-cuadrado para independencia de dos variables y la prueba exacta de Fisher.

Vamos a hacer en esta ocasión una prueba de la ji-cuadrado y dejaremos la prueba exacta de Fisher para otra ocasión.

Paso 2. Prueba de la ji-cuadrado de independencia

En primer lugar, vamos a hacerlo de la forma más sencilla, dejando que R nos construya la tabla de contingencia con las dos variables en estudio. Esto solo podremos hacerlo si tenemos cargado el conjunto de datos que queremos estudiar.

Abrimos el menú Estadísticos->Tablas de contingencia->Tabla de doble entrada… En la ventana emergente seleccionamos la variable de fila (survived) y la de columna (passengerClass). Acto seguido pulsamos en la pestaña Estadísticos y marcamos la opción "porcentaje por columnas". Esto lo hacemos para que nos diga, además de los números totales, el porcentaje de supervivencia en cada clase. Así podemos estudiar los resultados de las dos variables.

Como se muestra en la figura 4, en la parte inferior de esta ventana hay varias opciones más, que están relacionadas con la prueba de contraste. Vamos a marcar únicamente la opción "Test de

independencia de Chi-cuadrado" y pulsamos aceptar.

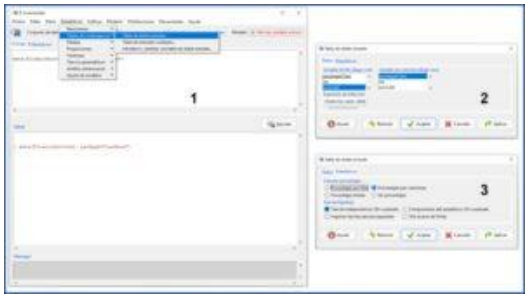

Figura 4. Tabla de contingencia y ji-cuadrado.

En la figura 5 podéis ver los resultados que se muestran en la ventana de salida.

```
Salida                                                    Ejecutar

Frequency table:
        passengerClass
survived 1st 2nd 3rd
      no  123 158 528
      yes 200 119 181

Column percentages:
        passengerClass
survived   1st  2nd   3rd
   no     38.1  57  74.5
   yes    61.9  43  25.5
   Total 100.0 100 100.0
   Count 323.0 277 709.0

        Pearson's Chi-squared test

data:  .Table
X-squared = 127.86, df = 2, p-value < 2.2e-16
```

Figura 5. Resultados de la prueba ji-cuadrado.

En primer lugar, tenemos la tabla de frecuencias absolutas. Podemos ver que sobreviven 200 de los 323 de primera clase, 119 de los 277 de segunda y 181 de los 709 de tercera. Aunque ya parece que hay diferencias, esto lo veremos mejor en la segunda tabla, que nos muestra los porcentajes por columnas.

Podemos ver que sobreviven el 61,9% de los pasajeros de primera clase, el 43% de los de segunda y el 25,5% de los de tercera. Ahora sí podemos ver las diferencias de forma clara, con un aumento de la probabilidad de sobrevivir en las clases más altas, pero ¿puede esta diferencia deberse a la casualidad?

Para contestar esta pregunta nos fijamos en la última línea de los resultados. Nos dice que el valor del estadístico ji-cuadrado es de 127.86, con 2 grados de libertad. Si la clase no influyese en la supervivencia, este valor debería ser próximo a 1. La probabilidad (el valor de p) de encontrar este valor o uno más alto por efecto del azar es de $2{,}2\times10^{-16}$, o sea, prácticamente 0 (y seguro que menor de 0,05).

La hipótesis nula de la prueba de la ji-cuadrado asume que las dos variables son independientes. Como la p es menor de 0,05, podemos rechazar la hipótesis nula y llegar a una conclusión importante: siempre que los datos sean representativos de la realidad, si alguna vez reflotan el Titanic y queremos viajar en él, mejor en primera clase. Por si acaso.

Paso 3. Introducción manual de la tabla de contingencia

Imaginad que no tenemos la base de datos, pero sabemos los resultados, ya sea en frecuencias absolutas o en porcentajes. En este caso, R no nos construirá la tabla de contingencia, pero sí que nos permite hacerlo de forma manual.

Seleccionamos la opción del menú Estadísticos->Tablas de contingencia->Introducir y analizar una variable de doble entrada… (figura 6). En la ventana emergente introducimos los nombres de las variables, marcamos el número de filas y columnas y rellenamos la tabla que se nos ofrece en la parte de debajo de la ventana.

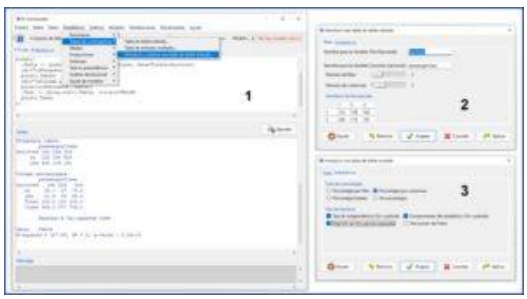

Figura 6. Entrada manual de la tabla de contingencia.

A continuación, pulsamos la pestaña Estadísticos y marcamos las opciones que nos interesen. Vamos a seleccionar la opción "Porcentajes por columnas" y a marcar todas las opciones del apartado "Test de hipótesis" excepto el test

exacto de Fisher que, como ya hemos dicho, vamos a dejar para otra ocasión.

En la figura 7 podéis ver la salida de resultados. En este caso le hemos pedido a R que nos muestre también los valores teóricos esperados si la hipótesis nula de independencia fuese cierta. Podemos ver cómo se desvían de los valores reales que hemos obtenido.

Figura 7. Resultados de la prueba ji-cuadrado.

Al igual que con el cálculo automático, ji-cuadrado tiene un valor de 127,86, con 2 grados de libertad, lo que supone un valor de p próximo a 0. No hay sorpresas, obtenemos el mismo resultado haciéndolo de las dos maneras.

Nos vamos…

Hemos visto en esta entrada cómo hacer la prueba de independencia de dos variables cualitativas utilizando la prueba de la ji-cuadrado.

Ya vimos en una entrada anterior que la prueba de la ji-cuadrado realiza una aproximación utilizando una distribución de probabilidad conocida, que no es otra que la distribución de la ji-cuadrado. Por otro lado, podemos utilizar una de las pruebas exactas, que calculan la probabilidad de forma directa, generando para ello todos los escenarios posibles en los que se produce la condición que queremos estudiar.

La prueba exacta para hacer este contraste de hipótesis es la denominada prueba exacta de Fisher que, aunque tiene unos requerimientos de cálculo más altos, debería ser la prueba de contraste de primera opción, especialmente con muestras pequeñas. Pero esa es otra historia…

Bibliografía

– Toledo E, Núñez-Córdoba M, Martínez-González MA. Datos categóricos y porcentajes: comparación de proporciones. En: Martínez-Sánchez MA, Sánchez-Villegas A, Toledo EA, Faulin J, eds. Bioestadística amigable, 3ª ed. Elsevier España, SL. Madrid, 2014; 147-74. (HTML)

– Fauquet J, Salafranca L. Pruebas no paramétricas y de libre distribución. En: Peró Cebollero M, Leiva Ureña D, Guardia Olmos J, Solanas Pérez A, eds. Estadística aplicada a las ciencias sociales mediante R y R-Commander. Ibergarceta Publicaciones SL, Madrid, 2012; 341-404. (HTML)

Correspondencia al autor

Manuel Molina Arias.
mma1961@gmail.com
Servicio de Gastroenterología.
Hospital Infantil Universitario La Paz.
Madrid. España.

Aceptado para el blog en octubre de 2021